PRO GRADU -TUTKIELMA

Juha Pyykkö

Malawilaisten lasten pituuskasvun mallintaminen sekamallein

TAMPEREEN YLIOPISTO
Informaatiotieteiden yksikkö
Tilastotiede
Heinäkuu 2014

Tampereen yliopisto
Informaatiotieteiden yksikkö
PYYKKÖ, JUHA: Malawilaisten lasten pituuskasvun mallintaminen sekamallein
Pro gradu -tutkielma, 41 s., 1 liites.
Tilastotiede
Heinäkuu 2014

Tiivistelmä

Tutkielmassa lasten pituuskasvun mallintamiseen malawilaisesta populaatiosta käytetään pitkittäisaineistoa LAIS-tutkimuksesta, jossa yksilöistä otettujen mittausten määrä ja mittausajankohdat vaihtelevat. Tällaisen aineiston mallinnukseen käytetään lineaarista sekamallia, jonka avulla havaintoaineiston epätasapainoisuus ei häiritse analyysia. Huomiota kiinnitetään myös mallin satunnaisvirheiden kovarianssirakenteeseen, jonka huomioon ottamalla mallinnusta voidaan tarkentaa. Mallintamalla pituutta iän muunnoksilla lineaarinen sekamalli tuottaa yksilöille sopivat mallit. Lisäksi mallinnusta ja aineistoa tarkastellaan pääkomponentti- ja ryhmittelyanalyysin avulla. Näitä menetelmiä hyödyntäen erilaisia kasvukäyriä on mahdollista havainnollistaa.

Asiasanat: lineaarinen sekamalli, pääkomponenttianalyysi, ryhmittelyanalyysi.

Sisältö

1 Johdanto

Ihmisen syntymäkoko on sikiökauden kasvun ja raskausajan pituuden yhteinen tulos (Cheung 2014). Pituuskasvun tekijöitä ovat niin perimä kuin elinolosuhteet. Lapsi kasvaa pituutta erityisesti ensimmäisen elinvuotensa aikana. Vanhempien pituus ei korreloi lapsen pituuden kanssa vielä syntymässä, vaan tulee esille vasta 2–3 vuoden iässä (Cameron & Bogin 2012). Pituuskasvu tasaantuu kolmen ikävuoden jälkeen ennen murrosiän kasvupyrähdystä, minkä jälkeen pituuskasvu jatkuu noin 20 ikävuoteen asti. Ihminen ei kasva lineaarisesti samaa suoraa, vaan kasvulla on erilaisia kulmakertoimia pituuskasvun aikana.

Saharan eteläpuolisessa Afrikassa lasten vajaaravitsemus on tunnettu ongelma, mikä vaikuttaa myös lasten pituuskasvuun. Vuonna 2012 Itä- ja Etelä-Afrikan alle viisivuotiaista lapsista 39 prosenttia oli kitukasvuisia (Unicef 2013). Kitukasvuiseksi määritellään, jos ikään suhteutetun pituuden z-arvo on alle -2. Z-arvot lasketaan Maailman terveysjärjestön (WHO) vertailupopulaatiosta, jolloin kitukasvuinen on itseisarvoltaan yli kahden keskihajonnan päässä vertailupopulaation mediaanista. Kasvuhäiriöiden syynä aliravitsemus on merkittävä tekijä. Aliravitsemuksessa on kyse erityisesti siitä, että lapsen paino ei ole pituuden mukainen. Vajaakehittynyt lapsi on silloin, kun hänen pituuskasvunsa jää jälkeen ikätasosta ravinnon puutteen vuoksi.

Tutkielman pituuskasvuaineisto on tutkimuksesta Lungwena Antenatal Intervention Study (LAIS), jossa mittaukset on tehty pitkittäistutkimuksena malawilaisille lapsille Lungwenan lähistöltä syntymästä viiden vuoden ikään asti. Aineistossa koko mittausajanjakson WHO:n iän mukaisen pituuden z-arvojen mediaani (keskiarvo) on –1.70 (–1.71), mikä tarkoittaa, että lungwenalaisten lasten pituuden mediaani sijoittuu koko WHO:n vertailupopulaation alimmaiseen viiteen prosenttiin. Näin aineisto tarjoaa mielenkiintoisen taustan pituuden mallintamiselle.

Pitkittäistutkimuksissa ei voida olettaa havaintojen riippumattomuutta, kun samalta osallistujalta mitataan useita havaintoja. Tällaisiin tilanteisiin soveltuvat sekamallit ottavat huomioon klusterien havaintojen riippuvuuden. Menetelmänä lineaarinen sekamalli mahdollistaa kiinteiden vaikutusten ja satunnaisvaikutusten mallintamisen. Näin koko aineistolle voidaan löytää yhteisiä tekijöitä, joiden lisäksi aineiston sisäisille klustereille voidaan muodostaa omia parametreja. Pitkittäistutkimuksessa luonnollisiksi klustereiksi valikoituvat yksittäiset lapset. Täten sekamallit tuottavat tutkimukseen osallistuneille lapsille omat mallit vertailtaviksi.

Mallinnuksien tarkoituksena on selittää lasten pituuskasvua lapsen iällä ja vertailla näitä kasvukäyriä. Lineaarinen sekamalli ei vaadi tasapainoista aineistoa, jolloin puuttuvat havainnot eivät vääristä tuloksia huomattavasti. Tutkielman aineistoon mukaan otetuilta lapsilta mittauskertojen mediaani on yksitoista ja mittauksia on 84 prosenttia suunnitellusta määrästä, josta puuttuva osuus on eri syistä epäonnistuneita mittauksia.

Kaiken kaikkiaan tutkielman tarkoituksena on mallintaa pituuskasvua populaatiolle ja yksilöille WHO:n standardien mukaan suhteellisen poikkeuksellisessa ai-

neistossa. Tutkielmaan ei sisälly syvempää analyysia olemassa olevista taustamuuttujista, koska tutkielman painotus on tilastollisessa mallintamisessa lääketieteellisten tutkimuskysymysten sijaan. Näin sukupuolten erottelu ja erityisesti interventioryhmien välisten eroavaisuuksien testaaminen ovat seikkoja, joista ollaan kiinnostuneita vasta LAIS-tutkimuksen alaisuudessa.

Tutkielman analyyseissa ja grafiikassa on käytetty ohjelmistoa R, versio 3.0.2 (R Core Team 2013). Ohjelmiston erityiset paketit ja funktiot mainitaan myöhemmin niitä koskevien menetelmien kohdalla.

2 Aineisto

2.1 LAIS

Tutkielman aineisto on osa Lungwena Antenatal Intervention Study -tutkimusta, jonka tarkoitus on selvittää raskaudenajan interventioiden vaikutusta ennenaikaisten syntymien vähentämiseen sekä äitien ja lasten hyvinvointiin synnytyksen jälkeen. Tutkimuksessa toimivat yhteistyössä Tampereen yliopiston lääketieteen yksikkö ja Malawin yliopiston College of Medicine. LAIS-tutkimus on rekisteröity sivustolle ClinicalTrials.gov tunnuksella NCT00131235. Tutkielman kannalta erityisessä tarkastelussa on lasten pituuskasvu, kun lapsia seurataan syntymästä 60 kuukauden ikään. Tässä luvussa on käytetty lähteenä Luntamon et al. (2012) artikkelia, jossa kuvataan tutkimuksen vaiheita.

Tutkimuslomakkeet on luotu tätä tutkimusta varten ja niiden tiedot on täytetty käsin terveyskeskuksissa Malawissa. Tämän jälkeen lomakkeiden pdf-kopiot on lähetetty Tampereen yliopiston kansainvälisen lääketieteen yksikköön. Omalta osaltani osallistuin aineiston keruuseen kesästä 2012 syksyyn 2013, kun siirsin tutkimusaineistoa kyselylomakkeista tietokantaan ja puhdistin aineistoa. Lisäksi vierailin Malawissa tutustumassa tutkimuksen käytäntöihin ja projektin muihin tutkimuksiin alkuvuodesta 2013.

2.1.1 Aineiston keräys

Tutkimusaineisto on kerätty Lungwenan maaseudulla, Malawissa, vuosien 2003 ja 2007 välillä. Tutkimukseen otettiin mukaan 1320 raskaana ollutta naista. Tutkimuksessa on kolme raskaudenaikaista interventioryhmää, joihin äidit satunnaistettiin tasaisten ja samankaltaisten ryhmien aikaansaamiseksi.

Tutkimuksessa käytetään lääkkeitä sulfadoksiini-pyrimetamiini ja atsitromysiini. Sulfadoksiini-pyrimetamiini on lääke malariaa vastaan. Atsitromysiini on antibiootti, jota käytetään bakteerien aiheuttamiin tulehduksiin muun muassa keuhkoissa, ja joka estää bakteerien lisääntymistä tulehduksista selviytymiseksi. Antibiootti tehoaa myös synnytyskanavan infektioihin.

Kontrolliryhmässä (436 naista) sulfadoksiini-pyrimetamiini on Saharan eteläpuoleisessa Afrikassa yleisen käytännön mukaan annettu kahdesti. Ensimmäisessä interventioryhmässä (SP-ryhmä, 441 naista) sulfadoksiini-pyrimetamiini on annettu raskauden aikana kuukausittain ja toisessa interventioryhmässä (AZI–SP-ryhmä, 442 naista) on annettu kaksi annosta atsitromysiinia kuukausittaisen sulfadoksiini-pyrimetamiinin lisäksi.

2.1.2 Aineiston puhdistus

Oleellisena osana tutkimusaineiston keruuta on aineiston täsmällisyys. LAIS-tutkimuksen kasvuaineisto kerättiin paperisille lomakkeille, joista tiedot siirrettiin tieto-

kantaan Cardiff TeleForm -ohjelmistoa käyttäen. Kyselylomake oli luotu jo tutkimuksen alkaessa, mutta sen TeleForm-pohjaa ei ollut olemassa. Osana harjoitteluani Tampereen yliopiston kansainvälisen lääketieteen yksikössä työnkuvaani kuului LAIS-aineiston siirtäminen pdf-tiedostoista tietokantaan, tehtävän organisointi ja aineiston puhdistaminen tutkimuskäyttöön.

Tein lomakkeille TeleForm-pohjat, joiden avulla lomakkeiden tiedot voitiin siirtää tietokantaan. Tämän jälkeen ryhmässämme toimi kahdeksan henkilöä syöttämässä tietoja lomakkeista tietokantaan. Toisin kuin suurissa, monien lomakkeiden tutkimuksissa, joissa osa lomakkeen kysymyksistä jätetään TeleForm-ohjelmiston automaattisesti ratkaistavaksi, neljän LAIS-lomakkeen jokainen kohta tarkistettiin aineiston syöttövaiheessa. Ihanteellisesti jokainen vastaus olisi ollut tämän jälkeen täsmällisesti oikein, mutta erilaisista inhimillisistä syistä, vastausten käsialan epäselvyydestä tai niiden kirjaamisesta väärään kohtaan lomaketta, aineistossa oli monia epäselvyyksiä.

Aineiston puhdistamisessa käytin hyväkseni muun muassa jakaumien reunimmaisten havaintojen ja päivämäärien loogisuuden tarkastelua sekä lasten kasvukäyriä ja mittausten z-arvoja poikkeavuuksien huomaamiseksi. Kaikki nämä löydökset olivat vielä aineiston siirtämistä tietokantaan, sillä näin korjasin vain tietokannan kirjaamisvirheitä vaikuttamatta lomakkeisiin kirjattuihin arvoihin.

Professori Per Ashornin kanssa kävimme läpi mittauksissa siten kirjattuja, mutta kasvukäyrältä huomattavasti poikkeavia havaintoja. Päädyimme poistamaan niistä osan, joita pidimme kirjausvirheinä. Muutama havainto korjattiin uskoen niiden olevan yhden merkin virheitä. Havainnot, jotka poikkesivat oletetulta kasvukäyrältä huomattavasti, poistettiin mittausvirheinä, jos niille ei löytynyt uskottavaa vaihtoehtoa. Viimeiseksi aineistoon lisättiin Luntamon aineiston mittaukset lapsen synnyttyä ja kuukauden iässä. Nämä muutokset sisältävä aineisto on niin sanottu puhdistettu aineisto.

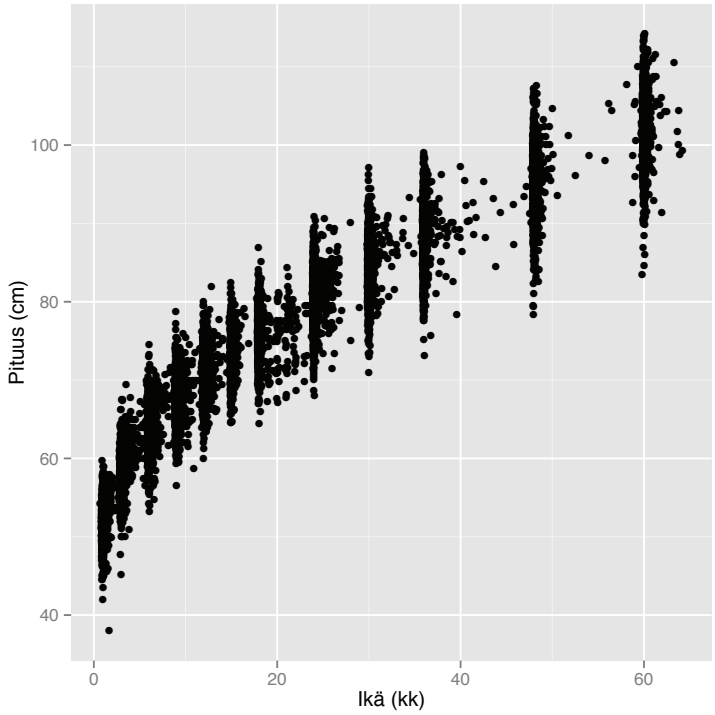
2.2 Mittaukset

Lapsilta on mitattu pituus selinmakuulla 24 kuukauteen asti, minkä jälkeen pituus on mitattu seisten. Lisäksi mittauksiin kuuluvat paino, yläkäsivarren ympärysmitta ja pään ympärysmitta. Muita tutkielmassa tiedossa olevia muuttujia ovat sukupuoli, kaksosuus ja interventioryhmä.

Tutkielman lopullinen aineisto koostuu lapsen iän mukaisten kuukausien 1, 3, 6, 9, 12, 15, 18, 24, 30, 36, 48 ja 60 suunnitelluista mittauksista. Tutkimuksen pituusmittauksia on yhteensä olemassa 12 suunniteltua mittauskertaa, sillä syntymästä (0 kk) on tiedossa vain paino. Jokaisella mittauskerralla lasten iällä on hajontaa, eivätkä kaikki lapset ole käyneet kaikilla vierailuilla, kun taas toiset lapset ovat käyneet suunnittelemattomilla vierailuilla. Lisäksi kaikki naiset lapsineen eivät lopulta aloittaneet tutkimuksessa ja osa osallistujista keskeytti tutkimukseen osallistumisensa.

Tutkielmassa käytetty on puhdistettua aineistoa, joka on päivitetty 28.11.2013. Tutkielmaan on kyseisestä aineistosta otettu mukaan ne lapsen pituuden sisältävät mittauskerrat, jolloin lapsen ikä on 0.50–65.00 kuukautta. Yhteensä käytettävissä on

12 453 pituushavaintoa 1230 lapselta. Kuviossa 2.1 on esitetty tutkielman aineiston pituushavainnot.



Kuvio 2.1. Tutkielman aineisto; lasten pituus ja ikä.

Mittauksista pystytään laskemaan myös WHO:n z-arvot lasten kasvulle, jolloin paino ja pituus voidaan suhteuttaa ikään erikseen ja keskenään. Z-arvot perustuvat normaalijakaumaan WHO:n vertailupopulaatiosta. Nämä z-arvot ovat hyviä mittareita sille, kuinka lasten kasvu kehittyy suureen populaatioon verrattuna. Z-arvot antavat erityisesti interventioryhmien vertailussa parempia vertailukohtia kuin pelkkä pituus ja paino. Iän mukainen paino *(weight-for-age z-score, WAZ)* ja pituus *(height-for-age z-score, HAZ)* ovat kaksi tutkimuksissa usein käytettyä mittaa. Tutkielman aineiston lungwenalaisten lasten koko mittausaikajakson kaikkien olemassaolevien havaintojen WAZ-keskiarvo (keskihajonta) on –1.04 (1.10) ja HAZ-keskiarvo (keskihajonta) on –1.71 (1.14). Näin tutkimuksen lapset ovat keskiarvoltaan yli yhden keskihajonnan päässä vertailupopulaation keskiarvosta, painoltaan pienimmässä 15 %:ssa ja pituudeltaan pienimmässä 5 %:ssa.

Tutkielmassa käytetään muuttujana ikää kuukausina, jotka on laskettu siten, että

$$\text{ikä} = \frac{\text{mittauspäivä} - \text{syntymäpäivä}}{365.25} \times 12 \text{ kk}.$$

3 Menetelmät

3.1 Lineaarinen sekamalli

Lineaarinen sekamalli *(linear mixed effects model, LME)* on tavallisen lineaarisen mallin laajennnus. LME:ssä lineaarisen mallin kiinteät vaikutukset saavat seurakseen klustereiden satunnaisvaikutukset. Tietyn klusterin sisällä havaintojen ajatellaan olevan riippuvia, kun taas klustereiden välillä havainnot oletetaan riippumattomiksi. Tutkielman aineistossa jokainen lapsi on klusteri ja sen havainnot ovat riippuvia. Mallinnuksessa aineisto jakautuu klusterin sisäiseen vaihteluun ja klustereiden väliseen vaihteluun.

Jiang (2007) sanoo esipuheessaan, että vaikutusten satunnaisuus on järkevää olettaa silloin, kun samalta yksilöltä on kerätty tietoa ajan mittaan. Näin havainnot ovat korreloituneita. LAIS-aineisto soveltuu lineaariselle sekamallinnukselle, jolloin satunnaiset vaikutukset ovat jokaiselle lapselle omat. Pituutta selitetään iällä ja sen eri muunnoksilla. Kiinteinä vaikutuksina voivat olla samat muuttujat kuin satunnaisvaikutuksina. Täten pituutta selitetään iällä koko populaatiossa ja jokaiselle yksittäiselle lapselle muodostetaan satunnaisista vaikutuksista omat lineaariset mallinsa. Nämä mallit sisältävät myös sukupuolen ja interventioryhmän vaikutukset, joita voidaan lopulta vertailla yleisellä tasolla.

Tutkielman aineiston mallintamiseen merkittävästi vaikuttava seikka on aineiston mittausten vaihteleva mittaushetki ja mittausten eri määrä osallistujittain. Fitzmaurice, Laird & Ware (2004, s. 188) nostavat juuri tämän seikan esille ja mainitsevat, että sekamallin satunnaisvaikutusten kovarianssirakenne ei vaadi samaa määrää havaintoja kultakin osallistujalta eikä sitä, että havainnot on mitattu samalla hetkellä. Näin epätasapainoinen pitkittäisaineisto soveltuu mallinnettavaksi sekamallein, mikä on erittäin tärkeä seikka tutkielman mallinnusongelmaa ratkottaessa.

Demidenko (2004) oikeuttaa tutkittavien satunnaisvaikutusten vakiokertoimen (b_0) mukaan ottamisen sekamallia käytettäessä sillä, että tutkittavilla voi samoista taustamuuttujien havaituista arvoista huolimatta olla eri alkutaso mitattavassa muuttujassa. Tässäkin tutkielman aineistossa, ja vauvoilla yleensä, raskauden pituus ja sen aikainen ravinto ja muut erilaiset tekijät tekevät vauvoista eri kokoisia syntymässä, vaikka ihmisen kasvu on palautettavissa ajan kulkuun. Näin satunnaisvaikutusten vakiokerroin on oleellinen osa tutkielman sekamallinnusta.

Tutkielmassa sovelletaan kahta R-ohjelmiston funktiota sekamallien estimoinnissa: Pinheiron & Batesin (2000) esittelemää funktiota `lme` paketissa `nlme` (Pinheiro et al. 2000) ja Batesin (2005) esittelemää funktiota `lmer` paketissa `lme4` (Bates et al. 2013).

3.1.1 Määrittely

Lairdin & Waren (1987) kehittämä sekamalli voidaan pitkittäiselle aineistolle ilmaista muun muassa Demidenkon (2004) mukaan

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n,$$

missä $\mathbf{y}_i$ ($j_i \times 1$) on i. yksilön vastevektori, $\mathbf{X}_i$ on i. yksilön kiinteiden vaikutusten suunnittelumatriisi, $\boldsymbol{\beta}$ ($m \times 1$) on kiinteiden vaikutusten vektori, $\mathbf{Z}_i$ on i. yksilön satunnaisten vaikutusten suunnittelumatriisi, $\mathbf{b}_i$ ($k \times 1$) on mallin satunnaisvaikutusten vektori ja $\boldsymbol{\epsilon}_i$ ($j_i \times 1$) on satunnaisvirheiden vektori.

Mikäli malliin sisältyy vakiotermejä, matriisien $\mathbf{X}_i$ ja $\mathbf{Z}_i$ ensimmäiset sarakkeet ovat ykkösvektoreita (merkitään $\mathbf{1}$). Oletuksena on, että $\mathbf{b}_i$ ja $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, ovat riippumattomia ja että

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i).$$

Myös klusterit oletetaan riippumattomiksi, mikä merkitään

$$\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{0}, \qquad i \neq j.$$

Havaintovektorin $\mathbf{y}_i$ kovarianssimatriisi on

$$\begin{aligned}
\mathbf{V}_i &= \mathrm{Cov}(\mathbf{y}_i) \\
&= \mathrm{Cov}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i) \\
&= \mathrm{Cov}(\mathbf{Z}_i\mathbf{b}_i) + \mathrm{Cov}(\boldsymbol{\epsilon}_i) \\
&= \mathbf{Z}_i\mathrm{Cov}(\mathbf{b}_i)\mathbf{Z}_i' + \mathbf{R}_i \\
&= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i. \qquad (3.1)
\end{aligned}$$

Näin marginaalijakauma $\mathbf{y}_i$ noudattaa moniuloitteista normaalijakaumaa

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i). \qquad (3.2)$$

Tässä marginaalijakaumassa osa $\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$ on klustereiden välistä vaihtelua ja osa $\mathbf{R}_i$ on klusterin sisäinen vaihtelu. Yleensä matriisi $\mathbf{D}$ ei ole tunnettu vaan sen estimointi on sekamallien keskeinen aihealue (Demidenko 2004). Kuten McCulloch & Searle (2001) tarkentavat, kiinteät vaikutukset ovat yhteydessä $\mathbf{y}_i$:n odotusarvoon, kun taas satunnaisvaikutusten mallimatriisi ja varianssi ainoastaan $\mathbf{y}_i$:n varianssiin. Jakaumaoletus (3.2) antaa perustan mallin parametrien suurimman uskottavuuden estimoimiselle (Nissinen 2009).

Näiden oletusten ollessa voimassa, kun $\mathbf{b}_i$ on annettu, niin

$$E(\mathbf{y}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i \qquad Cov(\mathbf{y}_i|\mathbf{b}_i) = \mathbf{R}_i.$$

3.1.2 Kiinteät vaikutukset ja satunnaisvaikutukset

Nissinen (2009) johtaa Hendersonin sekamalliyhtälön $\mathbf{y}$:n ja $\mathbf{b}$:n yhteisjakaumasta

$$\begin{aligned} f(\mathbf{y}, \mathbf{b}) &= f(\mathbf{y}|\mathbf{b})f(\mathbf{b}) \\ &= \frac{\exp\{-\frac{1}{2}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) + \mathbf{b}'\mathbf{D}^{-1}\mathbf{b}]\}}{(2\pi)^{(n+q)/2}|\mathbf{R}|^{1/2}|\mathbf{D}|^{1/2}}, \end{aligned}$$

missä $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$, $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_n')'$, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$, $\mathbf{Z} = diag\,(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, $\mathbf{R} = diag\,(\mathbf{R}_1, \ldots, \mathbf{R}_n)$ ja $\mathbf{D} = diag\,(\mathbf{D}, \ldots, \mathbf{D})$. Hendersonin (Herderson et al. 1959) sekamalliyhtälö on

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{3.3}$$

Sekamalliyhtälöstä (3.3) saadaan ratkaistua $\boldsymbol{\beta}$:lle paras lineaarinen ja harhaton estimaattori *(best linear unbiased estimator, BLUE)*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{3.4}$$

missä $\mathbf{V} = diag\,(\mathbf{V}_1, \ldots, \mathbf{V}_n)$, ja $\mathbf{b}_i$:lle paras lineaarinen ja harhaton ennustin *(best linear unbiased predictor, BLUP)*

$$\tilde{\mathbf{b}}_i = \mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}), \qquad i = 1, \ldots, n.$$

Matriisit $\mathbf{V}_i$ ja $\mathbf{Z}_i$ eivät ole tunnettuja vaan ne estimoidaan havaintoaineistosta, minkä jälkeen voidaan muodostaa empiirinen BLUP

$$\hat{\mathbf{b}}_i = \hat{\mathbf{D}}\mathbf{Z}_i'\hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}), \qquad i = 1, \ldots, n. \tag{3.5}$$

Sisäkorrelaatio on havaintojen korrelaatiokerroin klusterin sisällä ja se kertoo kuinka paljon aineiston vaihtelusta selittyy klustereiden välisellä vaihtelulla. Sisäkorrelaatio lasketaan

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2},$$

missä σ_b^2 on klusteritason varianssi ja σ_ϵ^2 on yksilötason varianssi. Mitä suurempi sisäkorrelaatio on, sitä enemmän klusterit selittävät aineiston vaihtelusta.

3.2 Estimointimenetelmät

Jotta BLUE (3.4) ja empiirinen BLUP (3.5) voidaan muodostaa, tarvitsee kovarianssimatriisin $\mathbf{V}_i$ (3.1) parametrit estimoida. Normaalijakaumaan perustuville sekamalleille on yleisesti kaksi iterointiin perustuvaa estimointitapaa: suurin uskottavuus *(maximum likelihood, ML)* ja rajoitettu suurin uskottavuus *(restricted/residual maximum likelihood, REML)*. ML on harhaisempi kuin REML ja McCulloch & Searle

(2001, s. 177–178) kommentoivat, että REML ei ole herkkä poikkeaville havainnoille. Yleisesti on suositeltua käyttää estimoinnissa REML-menetelmää ja Davison (2008, s. 660) suosittaa sitä erityisesti monimutkaisiin malleihin. Kuitenkin ML-menetelmää tarvitaan erilaisten kiinteiden vaikutusten vertailuun, sillä REML-mallien vertailu vaatii vertailtavilta malleilta samanlaisen mallin kiinteän osan. Tasapainoisen aineiston tapauksessa estimointimenetelmänä voidaan käyttää myös ANOVA-menetelmää, joka ei perustu iterointiin.

Searle, Casella & McCulloch (1992) ja Nissinen (2009) tarjoavat kattavan selosteen ML- ja REML-menetelmien teoriaan. ML ja REML ovat ominaisia osia sekamallimenetelmissä. Ohessa käydään läpi estimointimenetelmiä mainittuihin teoksiin pohjautuen.

3.2.1 Suurin uskottavuus

Estimaatit matriisien $\mathbf{D}$ ja $\mathbf{R}$ parametreille saadaan maksimoimalla log-uskottavuusfunktio

$$\log L_{\mathrm{ML}}(\mathbf{D}, \mathbf{R}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r},$$

missä $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, kun $\mathbf{b}$ ja $\boldsymbol{\epsilon}$ noudattavat normaalijakaumaa. Estimointi suoritetaan iteroimalla ja menetelmänä voidaan käyttää esimerkiksi Newton–Raphson-menetelmää, Fisherin scoring-menetelmää tai EM-algoritmia *(expectation and maximization)*.

3.2.2 Rajoitettu suurin uskottavuus

REML-menetelmässä maksimoidaan log-uskottavuusfunktio, jossa on tehty muunnos $\mathbf{z} = \mathbf{K}'\mathbf{y}$, missä $\mathbf{K}$ on sellainen täysiasteinen $n \times (n - p)$ -matriisi, että $\mathbf{K}'\mathbf{X} = \mathbf{0}$. Täten $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}'\mathbf{V}\mathbf{K})$, jolloin jakauma ei ole riippuvainen kiinteiden vaikutusten parametrivektorista $\boldsymbol{\beta}$. Näin mallin kiinteää osaa muutettaessa uskottavuusfunktio ei pysy samana. Tämän vuoksi REML ei sovellu kiinteältä osaltaan eroavien mallien vertailuun vaan sitä on käytettävä mallien vertailussa kiinteiden vaikutusten valinnan jälkeen.

REML-menetelmän tapauksessa estimaatit matriisien $\mathbf{D}$ ja $\mathbf{R}$ parametreille saadaan maksimoimalla uskottavuusfunktio

$$\log L_{\mathrm{REML}}(\mathbf{D}, \mathbf{R}) = -\frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r},$$

jonka ratkaisemiseksi voidaan käyttää samoja menetelmiä kuin ML-menetelmässä. Näistä menetelmistä Davison (2008, s. 659) mainitsee, että Newton–Raphson-menetelmästä voidaan johtaa RIGLS-algoritmi *(restricted iterative generalized least squares)*, kun taas EM-algoritmi on hitaampi mutta vakaampi.

3.3 Satunnaisvirheiden kovarianssirakenteet

Yksi osa sekamallinnusta on klusterin satunnaisvirheiden korrelaatioiden kovariansin mallintaminen. Klusterin satunnaisvirheiden korrelaatiot muodostuvat yksilölli-

sesti. Yksinkertaisin oletus on, että kovarianssimallissa (3.1) on diagonaalinen kovarianssirakenne, $\mathbf{R}_i = \sigma^2\mathbf{I}$, jossa yksilön havaintojen virheillä ei ole korrelaatiota.

Yksinkertaisimpana vaihtoehtoisena rakenteena tasakorrelaatiorakenteessa diagonaalin ulkopuoliset korrelaatiot (ϕ) ovat samoja koko kovarianssirakenteessa. Tällöin aikapisteiden välillä oletaan olevan sama korrelaatio riippumatta ajan etäisyydestä. Pinheiro & Bates (2000, s. 228) pitävät kuitenkin realistisempana, että havaintojen välinen korrelaatio heikkenee, kun mittauksien etäisyys kasvaa. Erityisesti pitkittäisaineistoissa korrelaation samana pysymisen olettaminen on huono valinta myös Fitzmauricen, Lairdin & Waren (2004, s. 168–169) mielestä.

Rakenteeton kovarianssirakenne ei tee oletuksia varianssista ja kovarianssista, vaan jokainen parametri muovautuu aineiston perusteella. Täten rakenne on tarkempi, mutta vaatii laskentatehoja enemmän kuin muut menetelmät. Tällaisessa tapauksessa havaintojen on oltava mitattu samoina mittausajankohtina.

Eksponentiaalinen kovarianssirakenne on ominaisin epätasapainoiselle aineistolle, kun autoregressiivinen kovarianssirakenne tarvitsee tasaisin väliajoin tehdyt mittaukset. Eksponentiaalinen kovarianssirakenne palautuu autoregressiiviseen AR(1)-kovarianssirakenteeseen, kun aikapisteet t ovat tasavälisiä. Nämä rakenteet kuitenkin olettavat, että korrelaatio hiipuu nopeasti nollaan. Lisäksi eksponentiaalinen kovarianssirakenne olettaa, ettei mittauksissa ole virheitä, jolloin samat tulokset saadaan toistettaessa mittaukset. Kovarianssirakenteen selkeyden, selittävyyden ja laskettavuuden kannalta satunnaisvirheiden eksponentiaalinen kovarianssirakenne on parhaiten perusteltavissa epätasapainoiselle aineistolle.

Eksponentiaalisessa kovarianssirakenteessa

$$\mathbf{R}_i = \sigma^2 \begin{pmatrix} 1 & \phi^{|t_{ij}-t_{ik}|} & \phi^{|t_{ij}-t_{ik}|} & \dots & \phi^{|t_{ij}-t_{ik}|} \\ \phi^{|t_{ij}-t_{ik}|} & 1 & \phi^{|t_{ij}-t_{ik}|} & \dots & \phi^{|t_{ij}-t_{ik}|} \\ \phi^{|t_{ij}-t_{ik}|} & \phi^{|t_{ij}-t_{ik}|} & 1 & \dots & \phi^{|t_{ij}-t_{ik}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{|t_{ij}-t_{ik}|} & \phi^{|t_{ij}-t_{ik}|} & \phi^{|t_{ij}-t_{ik}|} & \dots & 1 \end{pmatrix}$$

on yksilön i mittausten korrelaatiot ajanhetkillä $\{t_{i1}, \dots, t_{in}\}$.

Ottaen huomioon yllä mainitut kovarianssirakenteen oletukset on selvää, että kaikilla niillä on heikkoutensa, mutta niiden hyödyt kannustavat testaamaan niitä mallinnuksessa. Erityisesti epätasapainoisen pitkittäisaineiston tapauksessa eksponentiaalinen kovarianssirakenne ansaitsee huomiota mallin tarkasteluissa, sillä muut perinteisimmistä kovarianssirakenteista poikkeavat mallinnustavat olettavat tasapainoisen aineiston olemassaolon. Funktiossa `lme` objekti `corStruct` määrittää klusterin sisäisen korrelaatiorakenteen. Tätä ominaisuutta ei ole funktiossa `lmer`, minkä vuoksi tätä osiltaan parannettua funktiota ei voida käyttää, kun halutaan ottaa huomioon erilaiset kovarianssirakenteet.

Lisäksi on mainittava, että oletuksena näissä kovarianssirakenteissa varianssi on vakio. Tällöin satunnaisvirheiden varianssi pysyy samana koko mittausajanjakson. Varianssin muutosta voidaan testata ja mallintaa, jolloin homoskedastisesta klusterin sisäisestä varianssirakenteesta siirrytään heteroskedastiseen varianssien virheiden painotukseen. Funktiossa `lme` objekti `varFunc` määrittää klusterin sisäisen heteroskedastisuusrakenteen ja on oletuksena homoskedastinen.

3.4 Mallinvalinta

Aineistoon parhaiten soveltuvan mallin löytämiseksi voidaan asettaa erilaisia kriteerejä ja testejä. Mallin residuaalien normaalisuus ja vähäinen hajonta ovat visuaalisesti ja numeerisesti tarkasteltavia ominaisuuksia. Residuaalien tarkastelun lisäksi käytettävissä on mallien vertailuun ja mallien tuottamiin parametreihin perustuvia apukeinoja, joita esitellään seuraavassa.

3.4.1 Uskottavuussuhdetesti

Uskottavuussuhde *(likelihood ratio, LR)* muodostetaan kahden mallin vertailuun, jolloin

$$\delta = \frac{L(\hat{\boldsymbol{\Theta}}_0)}{L(\hat{\boldsymbol{\Theta}}_\alpha)}, \tag{3.6}$$

missä $\hat{\boldsymbol{\Theta}}_0$ on vähemmän parametrinen (rajoitettu) malli ja $\hat{\boldsymbol{\Theta}}_\alpha$ on vaihtoehtoinen (rajoittamamaton) malli. Oletusten vallitessa muunnos $-2 \log \delta$ noudattaa likimain χ^2_q-jakaumaa, jossa q on vertailtavien mallien parametrien lukumäärien erotus. Täten lause (3.6) muuttuu lauseeksi $-2 \log \delta = 2 \log L(\hat{\boldsymbol{\Theta}}_\alpha) - 2 \log L(\hat{\boldsymbol{\Theta}}_0)$, jonka arvoa verrataan χ^2_q-jakaumaan.

3.4.2 Informaatiokriteerit

Informaatiokriteereistä parhaiten tunnettuja ovat Akaiken informaatiokriteeri *(AIC)* sekä bayesilainen informaatiokriteeri *(BIC)*. Näissä informaatiokriteereissä logaritmoitu uskottavuusfunktio kerrotaan ja siihen lisätään rangaistustermejä. Niiden arvot lasketaan siten, että

$$\text{AIC} = -2\, l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}) + 2p \qquad \text{ja}$$
$$\text{BIC} = -2\, l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}) + p \ln(n),$$

joissa p on mallin parametrien määrä ja n on aineiston havaintojen määrä. Malleja vertailtaessa pienemmät informaatiokriteerit merkitsevät parempaa mallia.

Tähän saakka on esitelty lineaarisen sekamallin teoriaa. Seuraavat luvut käsittelevät pääkomponentti- ja ryhmittelyanalyysia, joiden avulla tilastollista mallinnusta voidaan havainnollistaa laajemmin.

3.5 Pääkomponenttianalyysi

Jos muuttujien määrä on suuri tai niiden yhdistelmät ovat vaikeasti tulkittavissa, muuttujien informaatiota voidaan haluta tiivistää. Toisaalta, jos muuttujien korrelaatiot ovat suuria, pääkomponenttianalyysi voi muokata näistä muuttujista yhden tai useamman muuttujan, jolloin useampi muuttuja voidaan muuttaa yhdeksi muuttujaksi. Näin aineistosta muodostetaan uusia muuttujia, jotka ovat alkuperäisten muuttujien lineaarikombinaatioita. Luvun lähteenä on käytetty Johnsonin & Wichernin teosta (2007).

3.5.1 Pääkomponenttien määrittäminen

Pääkomponentit määritellään seuraavasti. Olkoon $\mathbf{U} = (U_1, U_2, \ldots, U_p)'$ p muuttujan vektori. Muuttujat voidaan halutessa standardoida muuttujien mittojen yhtenäistämiseksi. Näiden muuttujien kovarianssimatriisi on $\mathbf{\Sigma}$ $(p \times p)$. Näistä muuttujista muodostetaan pääkomponentit

$$\begin{aligned}
T_1 &= \mathbf{a}_1'\mathbf{U} = a_{11}U_1 + a_{12}U_2 + \cdots + a_{1p}U_p \\
T_2 &= \mathbf{a}_2'\mathbf{U} = a_{21}U_1 + a_{22}U_2 + \cdots + a_{2p}U_p \\
&\vdots \\
T_p &= \mathbf{a}_p'\mathbf{U} = a_{p1}U_1 + a_{p2}U_2 + \cdots + a_{pp}U_p,
\end{aligned}$$

missä $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p)$ on kerroinvektori muuttujille. Pääkomponentit $\mathbf{T}$ muodostetaan siten, että ensimmäisen pääkomponentin, T_1, lineaarikombinaatio $\mathbf{a}_1'\mathbf{U}$ maksimoi varianssin $\mathrm{Var}(T_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$ ehdolla $\mathbf{a}_1'\mathbf{a}_1 = 1$. Tämän jälkeen k. pääkomponentti T_k muodostetaan siten, että maksimoidaan $\mathrm{Var}(T_k)$ ehdoilla $\mathbf{a}_k'\mathbf{a}_k = 1$ ja $\mathrm{Cov}(T_k, T_j) = 0$, missä $k > 1$ ja $k > j$. Täten pääkomponenteiksi muodostuu rajoitetulla kerroinvektorin pituudella suurimman vaihtelun sisältäviä lineaarikombinaatioita, jotka ovat korreloimattomia keskenään. Ensimmäiseen pääkomponenttiin sisältyy suurin aineiston sisältämä vaihtelu.

Pääkomponentit voidaan muodostaa niin, että muuttujien $\mathbf{U}$ kovarianssimatriisille $\mathbf{\Sigma}$ tehdään ominaisarvohajotelma $\mathbf{\Sigma} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$, missä $\mathbf{P}$ on ortogonaalimatriisi ($\mathbf{P}\mathbf{P}' = \mathbf{P}'\mathbf{P} = \mathbf{1}$) ja $\mathbf{\Lambda}$ on diagonaalimatriisi elementein $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Matriisin $\mathbf{\Lambda}$ diagonaalielementit ovat $\mathbf{\Sigma}$:n ominaisarvoja, jolloin i. pääkomponentin varianssi on λ_i. Tällöin k ensimmäistä pääkomponenttia sisältävät kokonaisvaihtelusta $100\ \% \times (\sum_{i=1}^{k} \lambda_i)/(\sum_{i=1}^{p} \lambda_i)$. Lopulta pääkomponenttien lukumäärän valitseminen voidaan perustella niiden tarpeeksi suurella selitysosuudella kokonaisvaihtelusta.

Näin muodostetaan yhdistelmiä, jotka parhaiten kuvaavat aineiston varianssia. Ensimmäistä pääkomponenttia voidaan kuvailla aineistoon parhaiten asettuvana suorana, kun toinen pääkomponentti on ensimmäisen pääkomponentin virheisiin parhaiten asettuva suora. Seuraavat pääkomponentit ovat suoria, jotka asettuvat parhaiten edellisten pääkomponenttien yhteisvirheisiin.

Vaikka ensimmäiseen pääkomponenttiin sisältyy aineiston suurin vaihtelu, muut pääkomponentit tuovat esille toisenlaisia riippuvuussuhteita. Toisinaan tällaiset riippuvuussuhteet ovat alkuperäisessä aineistossa huomaamattomia, jolloin pääkomponenttien avulla niitä voidaan havainnoida ja tutkia.

3.5.2 Pääkomponentit mallinnuksessa

Mikäli sekamallin satunnaisvaikutuksista määritetään pääkomponentit, voidaan niitä käyttää parametrien approksimointiin mallissa. Kun satunnaisvektorin $\mathbf{b}$ varians-

simatriisista muodostettu diagonaalimatriisi on

$$\mathbf{D} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix},$$

niin standardoitujen muuttujien pääkomponenttivektori on $\mathbf{u} = \mathbf{A}'\mathbf{D}^{-\frac{1}{2}}\mathbf{b}$. Tästä voidaan johtaa

$$\begin{aligned} \mathbf{u} &= \mathbf{A}'\mathbf{D}^{-\frac{1}{2}}\mathbf{b} \\ \mathbf{A}\mathbf{u} &= \mathbf{D}^{-\frac{1}{2}}\mathbf{b} \\ \mathbf{b} &= \mathbf{D}^{\frac{1}{2}}\mathbf{A}\mathbf{u} \\ &= \mathbf{D}^{\frac{1}{2}}\left(u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + \cdots + u_p\mathbf{a}_p\right), \end{aligned}$$

mistä approksimaatio ensimmäistä pääkomponenttia käyttäen on $u_1\mathbf{D}^{\frac{1}{2}}\mathbf{a}_1$. Jos pääkomponentit halutaan palauttaa alkuperäisten muuttujien kombinaatioksi standardoidun aineiston tilanteessa ensimmäisen pääkomponentin osalta, voidaan muodostaa $\mathbf{y}^* = \mathbf{Z}\mathbf{b} \approx u_1\mathbf{Z}\mathbf{D}^{\frac{1}{2}}\mathbf{a}_1$.

3.6 Ryhmittelyanalyysi

Ryhmittelyanalyysissa *(cluster analysis)* havainnot tai muuttujat pyritään jakamaan ryhmiin niiden arvojen perusteella. Ryhmien ominaisuuksia tai määrää ei määrätä ennakkoon. Ryhmittely perustuu havaintojen väliseen etäisyyteen ja havaintojen läheisyys yhdistää ryhmien jäseniä. Tutkielmassa perehdytään havaintoja ryhmitteleviin menetelmiin jatkuville muuttujille. Näistä menetelmistä on kuitenkin huomattava, että ryhmittelyanalyysi on varsin heuristinen menetelmä ja siten sen tuottamat tulkinnat ovat subjektiivisia. Koska ryhmittelyanalyysi ei tuota testituloksia, tuotettujen ryhmien tulkinta on tutkijan vastuulla.

Ryhmittelyanalyysimenetelmistä mainitaan hierarkkisen ryhmittelyanalyysin neljä variaatiota. Näissä menetelmissä ei tehdä oletusta taustalla olevasta jakaumasta. Havainnnot mahdollisesti standardoidaan ennen analyysia ja usein lasketaan havaintojen välinen *euklidinen etäisyys*

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} \sqrt{(x_{ij} - x_{i'j})^2},$$

missä yksilön i havaintoa muuttujassa j verrataan toisiin. Muita jatkuvien muuttujien etäisyysmittoja ovat muun muassa *itsearvoetäisyys*, *maksiminormietäisyys*, *Mahalanobis-etäisyys* ja *Minkowski-etäisyys*.

Lisäksi on mainittava ei-hierarkkinen K-keskiarvon *(K-means)* ryhmittelyanalyysi, jossa on aluksi asetettava ryhmien määrä *(k)* valitsemalla ryhmille keskipisteet. Tällöin menetelmässä havaintoja siirretään yksi kerrallaan ryhmään, jonka keskipisteeseen havainnoilla on pienin etäisyys. Jokaisen siirron jälkeen muutettujen ryhmien keskipisteet lasketaan uudestaan ja siirtelyä toistetaan, kunnes ryhmittelyssä ei enää tapahdu muutoksia. K-keskiarvon ryhmittelyn tulokset saattavat riippua asetetuista alkuarvoista, joten niin eri alkuarvojen asetus kuin oletettujen ryhmien lukumäärä vaativat tulosten vertailua parhaiden ryhmien löytämiseksi. Keskipisteiden valinnassa voidaan käyttää esimerkiksi hierarkkisen ryhmittelyn tuottamia ryhmäkeskipisteitä, äärimmäisiä keskipisteitä tai ennakkotietoon perustuvia keskipisteitä.

3.6.1 Hierarkkinen ryhmittely

Hierarkkisen ryhmittelyn menetelmässä jokainen havainto muodostaa aluksi oman ryhmänsä. Sen jälkeen toisiaan lähimpänä olevat havainnot yhdistetään uudeksi ryhmäksi, jolloin ryhmien lukumäärä pienenee yhdellä. Tämän jälkeen etsitään taas havainnot, jotka ovat lähimpänä toisiaan ja ne yhdistetään samaan ryhmään. Havainto voi olla lähimpänä toista yksittäisen tai useamman havainnon ryhmää. Tätä voidaan toistaa kunnes ryhmiä on enää kaksi tai jokin muu ryhmien määrä päätetään, kun valitun kriteerin arvo on halutun kaltainen. Myös ryhmittelymenetelmiä käytettäessä on syytä standardoida käytettävä aineisto, jotta suuremman mitta-asteikon muuttujat eivät vaikuta etäisyysmittaan enemmän kuin pienempiä havaintoarvoja saavat muuttujat, mikäli näillä muuttujilla ei haluta olevan ylimääräistä vaikutusta mitta-asteikkonsa vuoksi.

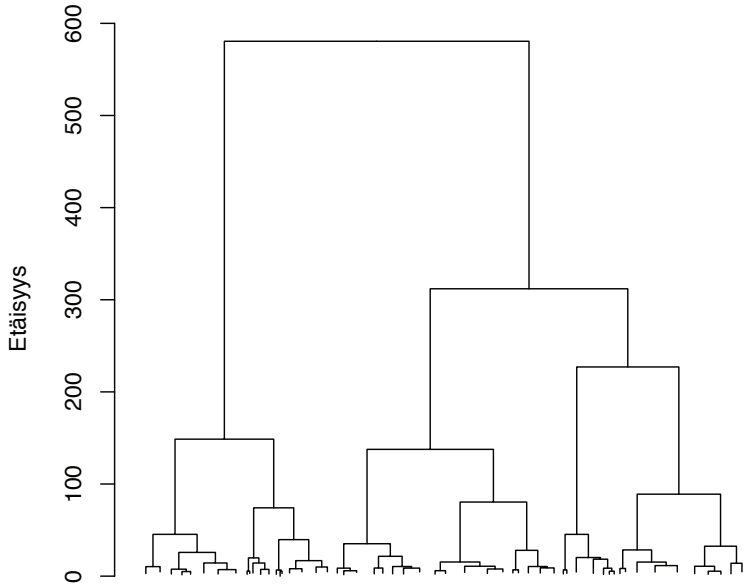
Hierarkkisessa ryhmittelyssä voidaan käyttää erilaisia ryhmittelymenetelmiä, kuten yhden sidoksen menetelmää *(single linkage)*, jossa kahden ryhmän etäisyydeksi määritellään pienin etäisyys kahden eri ryhmään kuuluvan havainnon välillä. Tällä menetelmällä on taipumus yhdistää ryhmiä hyvin pienellä kynnyksellä, jolloin ryhmät niin sanotusti ketjuuntuvat, kun juuri yhdistettyyn ryhmään yhdistetään taas uusi havainto (Hastie, Thibshirani & Friedman 2008, s. 525).

Kun ryhmien väliseksi etäisyydeksi määritellään vastaavasti suurin etäisyys, on kyse täydellisen sidoksen menetelmästä *(complete linkage)*. Näin uhkana voi kuitenkin olla se, että jotkin havainnot ovat lähempänä jotain toista ryhmää kuin omaansa (Hastie, Thibshirani & Friedman 2008, s. 524). Keskimääräisen sidoksen menetelmä *(average linkage)* laskee keskiarvon ryhmien välisten havaintoparien etäisyyksistä. Nämä kolme mainittua menetelmää ovat havainnollisestettu graafisesti Johnsonin & Wichernin teoksessa (2007, s. 681). Neljäntenä mainittavana menetelmänä Wardin menetelmässä pyritään maksimoimaan ryhmien homogeenisuus, jolloin yhdistetään ne ryhmät, joiden sisäistä vaihtelua mittaava neliösumma kasvaa vähiten.

Kun ryhmittely aloitetaan kaikki havainnot omana ryhmänään ja ryhmien yhdistelyä jatketaan, on huomattava, että havaintoja ei enää siirretä ryhmästään pois. Jos ryhmittelyn halutaan sopeutuvan ryhmittelyn kuluessa, tällöin voidaan käyttää K-keskiarvon menetelmää. Ryhmien alkuarvoiksi voidaan asettaa hierarkkisella menetelmällä saadut ryhmäkeskiarvot, jotta klusterit ryhmittyvät uudelleen.

3.6.2 Etäisyys, samankaltaisuus ja puukuvio

Ryhmittelyn tuloksia on selkeää tulkita puukuvion *(dendrogram)* avulla. Siinä ryhmät jaotellaan etäisyyden tai samankaltaisuuden mukaan eri haaroihin hierarkkisen ryhmittelyn perusteella. Kuviossa 3.1 on esitetty esimerkki puukuviosta tutkielman aineiston tapauksessa, kun jokainen 1230 yksilöstä on yhdistetty lopulta yhteen ryhmään. Haluttu määrä ryhmiä voidaan muodostaa katkaisemalla ryhmittely valitulta kohtaa y-akselia. Alhaalta ylöspäin yksilöitä on koottu ryhmiksi pienimmän etäisyyden mukaan. Vaakatason mukaiset suorat kuvaavat ryhmäjakoa ja pystyviivat merkitsevät ryhmien etäisyyttä. Siten mitä korkeammalla ryhmät yhdistyvät, sitä enemmän näiden ryhmien välillä on etäisyyttä.



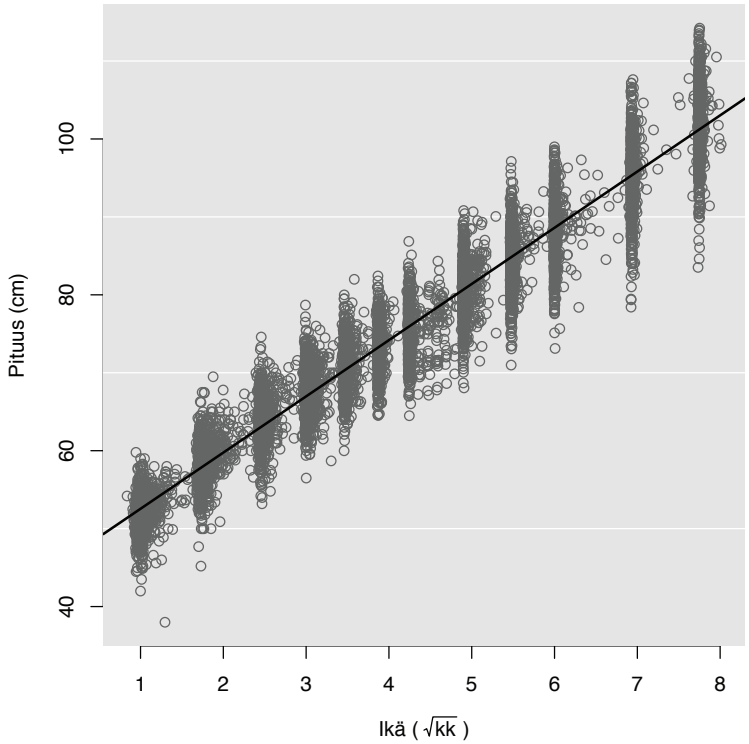
Kuvio 3.1. Puukuvio yksilöiden ryhmittelystä satunnaisvaikutusten perusteella Wardin menetelmällä, kun havaintomäärän selkeyttämiseksi haarat on katkaistu etäisyydessä 5.0. Tällöin muodostuu 55 ryhmää.

4 Analyysi

4.1 Mallinvalinta

Tutkielmassa lapsen pituutta selitetään lapsen iällä noin viiteen ikävuoteen asti. Vaihtoehtoisissa menetelmissä voi olla soveltuvaa muokata selitettävää muuttujaa, mutta lineaarisessa sekamallissa on helppoa käyttää selittävän muuttujan muunnoksia. Koska lapsen pituuskasvua ei voida kuvata suorana viivana iän suhteen, soveltuvien iän muunnosten löytäminen on välttämätöntä tutkielman pituuskasvumallinnuksissa.

Aineistossa erityisen mielenkiintoisen muutoksen tuo jo iän neliöjuurimuunnos, jonka avulla pituuskasvu kuviossa 4.1 on jo huomattavasti lineaarisempaa kuin alkuperäinen aineisto kuviossa 2.1.



Kuvio 4.1. Iän neliöjuurimuunnos ja lineaarisen mallin (4.1) regressiosuora.

Kuvioon 4.1 on myös sovitettu lineaarinen regressiosuora

$$\text{pituus} = 45.35 + 7.21\,\sqrt{\text{ikä}}, \tag{4.1}$$

jossa lineaarisen mallin oletukset eivät kuitenkaan ole kunnossa, koska aineisto ei ole riippumaton. Silti esimerkki antaa kuvauksen aineiston muutoksen tuomasta hyödystä. Pelkkä kiinteiden vaikutusten lineaarinen malli (4.1) ei siten pysty täydelliseen mallinnukseen, ja lisäksi sen residuaaleissa ilmenee huojuntaa iän vaihdellessa.

Klusterit, eli aineiston tapauksessa yksittäiset lapset, pakottavat sekamallin käyttöön. Neliöjuurimallinnus ei kuitenkaan pelkästään pysty aivan tyydyttävään mallinnustulokseen edes sekamallina, sillä havaintoaineisto aaltoilee regressiosuoran ympärillä. Niinpä mallinnusta on jatkettava lisäämällä malliin parametreja.

Jos iän neliöjuurimuunnosta pidetään iän muunnosten perustana, mallinnuksen parantamiseksi voidaan malliin lisätä iän potenssin puolikkaita. Pienempien potenssien lisäysten sisältäminen malliin vaikeuttaa mallin tulkintaa, kun tasapainoillaan ymmärrettävyyden ja selitettävyyden välillä. Lisättäessä sekamalliin iän neliöjuuren eri potensseja parhaaksi malliksi valikoituu

$$\begin{aligned}\text{pituus}_i = {} & (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\sqrt{\text{ikä}_i} \\ & + (\beta_2 + b_{2i})\,\text{ikä}_i + (\beta_3 + b_{3i})\,\text{ikä}_i^{\frac{3}{2}} + \epsilon_i,\end{aligned} \tag{4.2}$$

missä i on yksilö aineistosta. Iän muunnosten lisääminen niin kiinteisiin vaikutuksiin kuin satunnaisvaikutuksiin on perusteltu mallinvalinnan tunnuslukujen ja residuaalien jakaumien vertailulla eri mallien kesken.

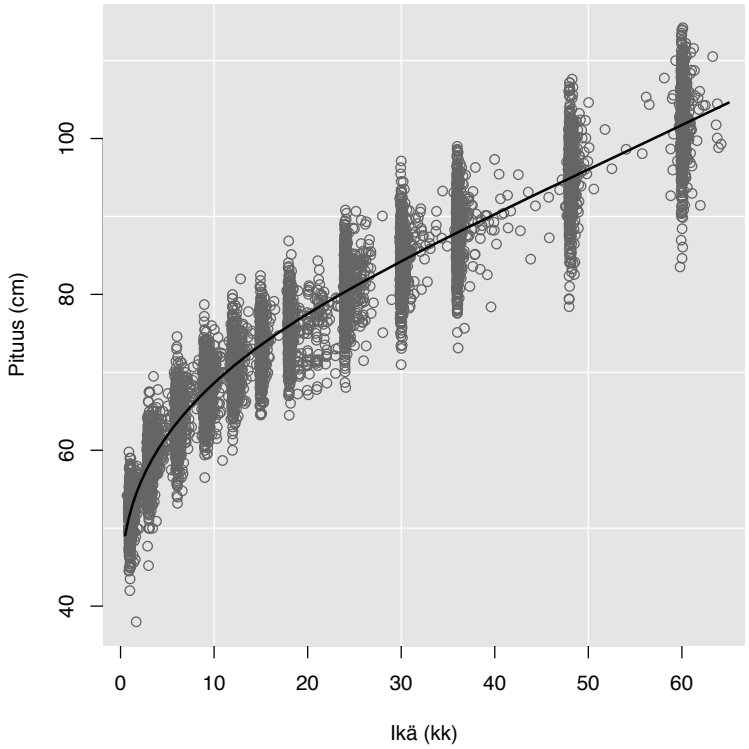
Kokonaisuudessaan mallinvalinta perustuu informaatiokriteereiden, uskottavuussuhdetestien tuloksien ja residuaalien vertailuun, kun malliin lisätään niin kiinteitä vaikutuksia kuin satunnaisvaikutuksia mallia kasvattaen. Ikä ja ikä potenssiin $\frac{3}{2}$ tarkentavat mallia hyödyttävästi, kun taas iän neliömuunnos ei enää paranna mallia riittävästi.

Mallista on tarkastettava myös sen klustereiden satunnaisvirheiden kovarianssirakenne. Funktion `nlme` oletuksena kovarianssien rakenteeksi on $\sigma^2\mathbf{I}$, jonka uskottavuutta testattiin mallin kiinteiden vaikutusten ja satunnaisvaikutusten valinnan jälkeen. Sekamallin (4.2) kovarianssirakenne on lopulta eksponentiaalinen, minkä valinnasta raportoidaan tarkemmin luvussa 4.2. Eri mallien vertailu uusittiin tämän kovarianssirakenteen valinnan jälkeen, mutta malliin valikoituivat samat parametrit kuin kovarianssirakenteen $\sigma^2\mathbf{I}$ tapauksessa.

Tutkielman aineiston tapauksessa varianssin heteroskedastisuudelle ei ole riittäviä perusteita, joten kovariassirakenteen mallinnus riittää mallin tapauksessa, kun malliin valitut parametrit tasoittavat residuaalien jakaumaa. Kun mallin (4.2) kovarianssirakenne on eksponentiaalinen, mallin estimoidut parametrit ovat

$$\begin{aligned} \hat{\beta}_0 &= 42.117 & \hat{\sigma}_{b_0} &= 2.018 \\ \hat{\beta}_1 &= 10.455 & \hat{\sigma}_{b_1} &= 1.910 \\ \hat{\beta}_2 &= -0.861 & \hat{\sigma}_{b_2} &= 0.551 \\ \hat{\beta}_3 &= 0.065 & \hat{\sigma}_{b_3} &= 0.042 \\ & & \hat{\sigma}_{\epsilon} &= 1.402. \end{aligned}$$

Kuviossa 4.2 sekamalli (4.2) on kuvattu ilman satunnaisvaikutuksia ($b_{ji} = 0$; $j = 0, 1, 2$; $i = 1, \ldots, n$). Yksittäisiä lapsia klustereina pitävän sekä kiinteitä vaikutuksia ja satunnaisvaikutuksia sisältävän sekamallin (4.2) residuaalien kvantiili-kvantiili–kuviossa 4.3 nähdään residuaalien jakauman huomattava huipukkuus. Sekamallin (4.2) residuaalien varianssi on 1.423. Mallin soveltuessa hyvin yksittäisiin



Kuvio 4.2. Sekamallin (4.2) kiinteiden vaikutusten käyrän sijoittaminen aineistoon.

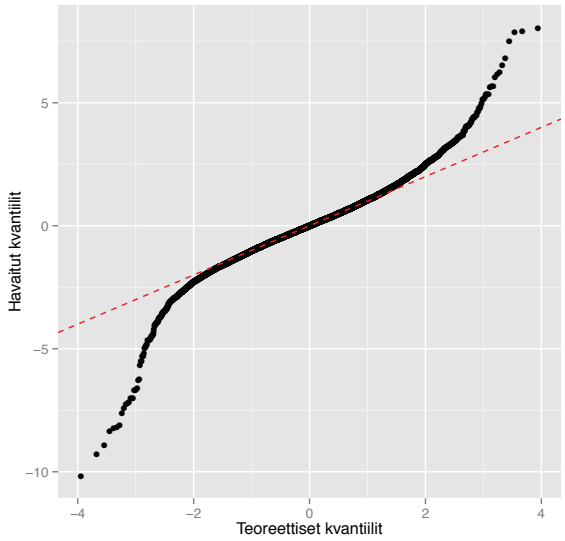
klustereihin on oletettavaa, että klustereiden suuri määrä verrattuna klusterin mittauksiin tuottaa vähän hajontaa. Mallin sisäkorrelaatio $\rho = 0.803$, mikä selittyy iän mukaisten pituuksien laajalla hajonnalla. Se myös vahvistaa ajatusta sekamallien toimivuudesta yksilöiltä kerätyn pitkittäisaineiston mallinnuksessa.

Valitun sekamallin (4.2) satunnaisvaikutuksilla on huomattavaa korrelaatiota, mikä on esitetty kuviossa 4.4. Tässä huomataan myös ikä-muuttujan satunnaisparametrin negatiivinen korrelaatio iän muihin muunnoksiin verrattuna, mikä vaikeuttaa mallin tulkintaa. Luvussa 4.4 satunnaisvaikutusten tulkitsemista kohennetaan pääkomponenttianalyysin avulla.

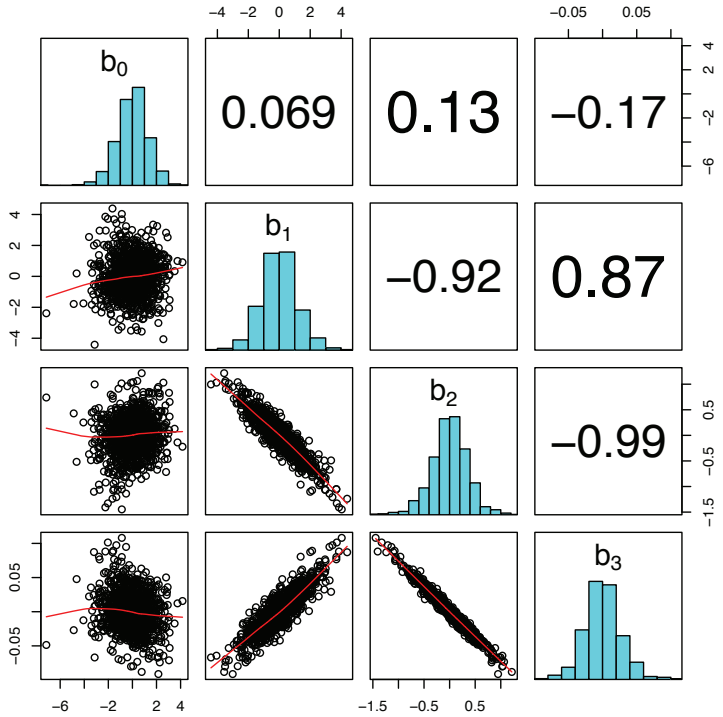
4.2 Kovarianssirakenne

Aineistoa mallinnettiin satunnaisvirheiden eksponentiaalisen kovarianssirakenteen kanssa mallinvalinnan kriteereiden vertailun jälkeen. Näin myös residuaalien jakauman huipun vinoutta saatiin hieman korjattua verrattuna tasakorrelaatiorakenteeseen, jossa havainnoille ei oleteta korrelaatiota. Tämä on esitetty kuviossa 4.5, kun on käytetty Pearsonin residuaaleja, $\boldsymbol{r}_p = (\boldsymbol{y} - \hat{\mu})/\sqrt{\text{var}(\hat{\mu})}$. Poikkeavan kovarianssirakenteen huomioiminen pakottaa funktion `lme` käyttöön, sillä funktiolla `lmer` ei voi määrittää satunnaisvirheiden kovarianssirakenteen muotoa.

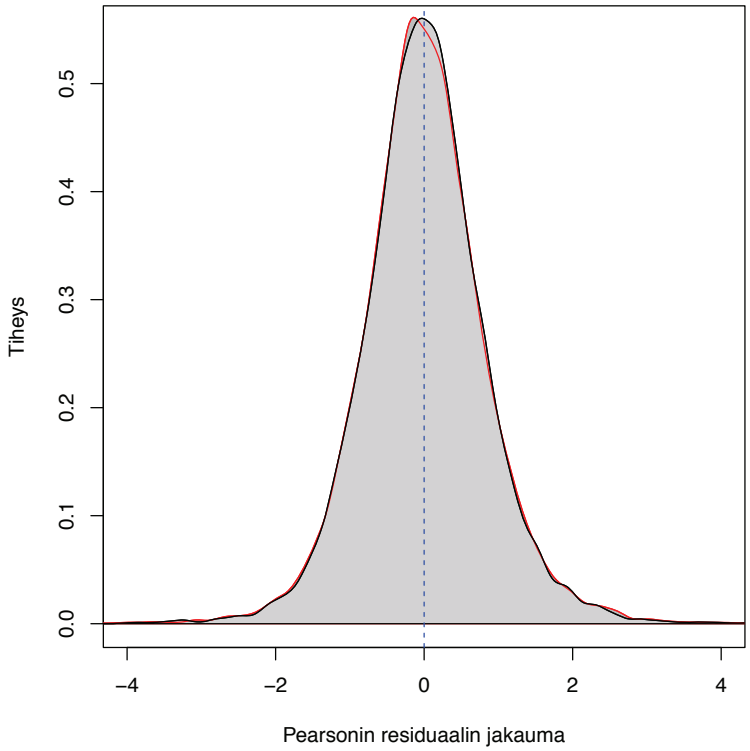
Pitkittäisaineiston kannalta on hyvä käyttää jonkinlaista satunnaisvirheiden korrelaatiorakennetta, joka tuo esiin klustereiden riippuvuuden. Tämä on sekamallin



Kuvio 4.3. Sekamallin (4.2) residuaalien kvantiili-kvantiili–kuvio eksponentiaalisella kovarianssirakenteella. Katkoviiva kulkee ensimmäisen ja kolmannen kvartiilin kautta.



Kuvio 4.4. Sekamallin (4.2) satunnaisvaikutusten väliset korrelaatiot korrelaatiokertoimina oikeassa kulmassa ja pisteparvina vasemmassa kulmassa. Diagonaalisella akselilla satunnaisvaikutusten itsenäiset jakaumat.



Kuvio 4.5. Sekamallin (4.2) residuaalien (r_p) tiheysfunktiot tasakorrelaatiorakenteella *(punainen)* ja eksponentiaalisella kovarianssirakenteella *(musta)*.

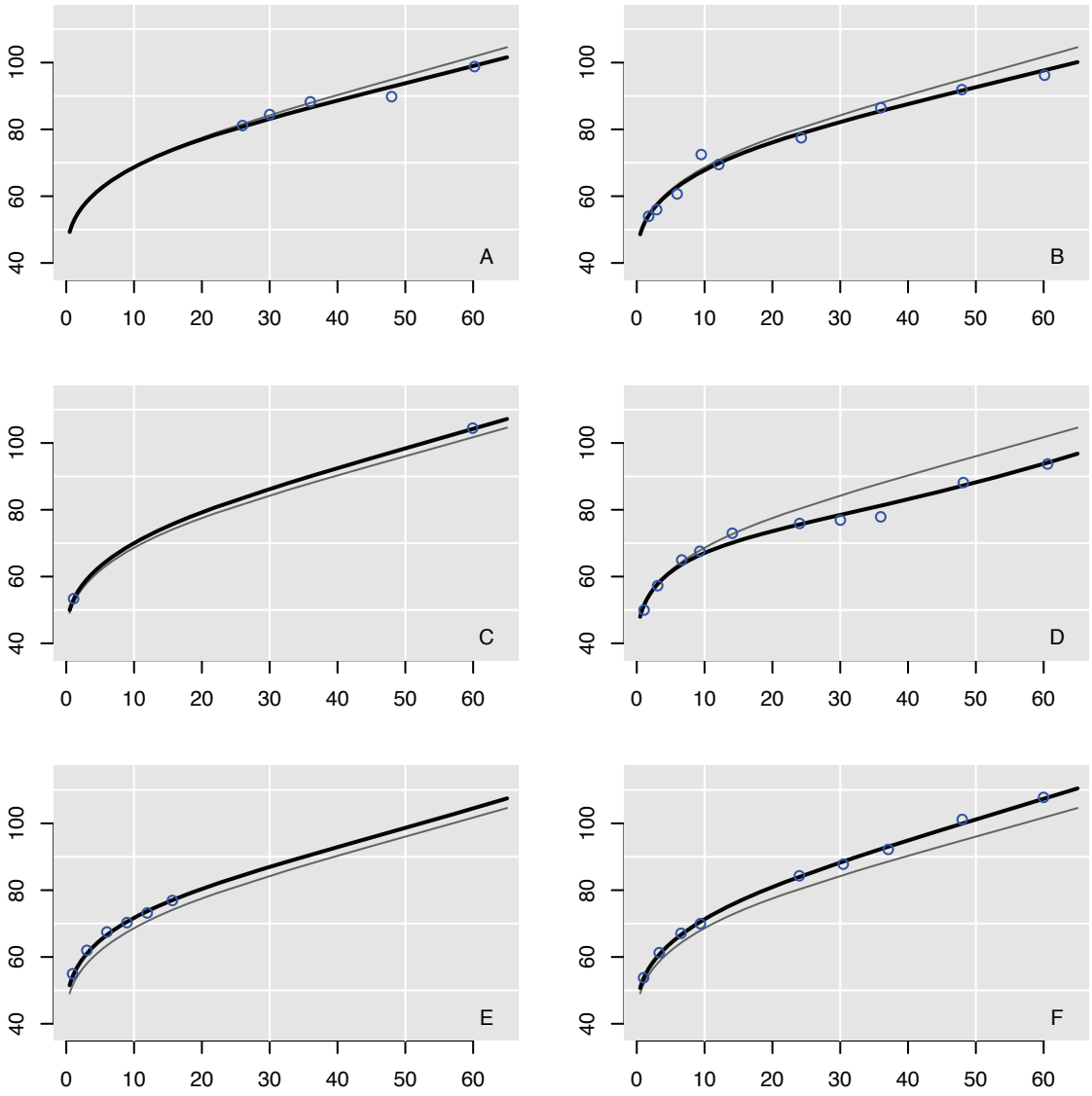
erityinen hyvyys, joka mahdollistaa mallinnuksen tavallista regressiomallia mukaillen, mutta antaa epätasapainoisille klustereille mallinnusmahdollisuuden.

4.3 Mallin toimivuus

Valittu sekamalli tuottaa populaatiolle ja yksilöille lineaariset regressiokäyrät. Se voidaan estimoida mittausvirheistä ja puuttuvista havainnoista huolimatta. Mallin tulkintaa ja kasvukäyrien ryhmittelyä käsitellään vielä seuraavissa luvuissa, kun tässä luvussa tarkastellaan mallin sopivuutta mallin antamilla mahdollisuuksilla.

On huomautettava, että syitä puuttuville havainnoille ei ole eritelty tutkielman aineistossa. Havainnon puuttumisen syynä voi olla niin paikkakunnalta pois muuttaminen kuin yksilön kuolema. Myöskään sairauksia tai kuoleman syytä ei pystytä huomioimaan tutkielman aineistossa.

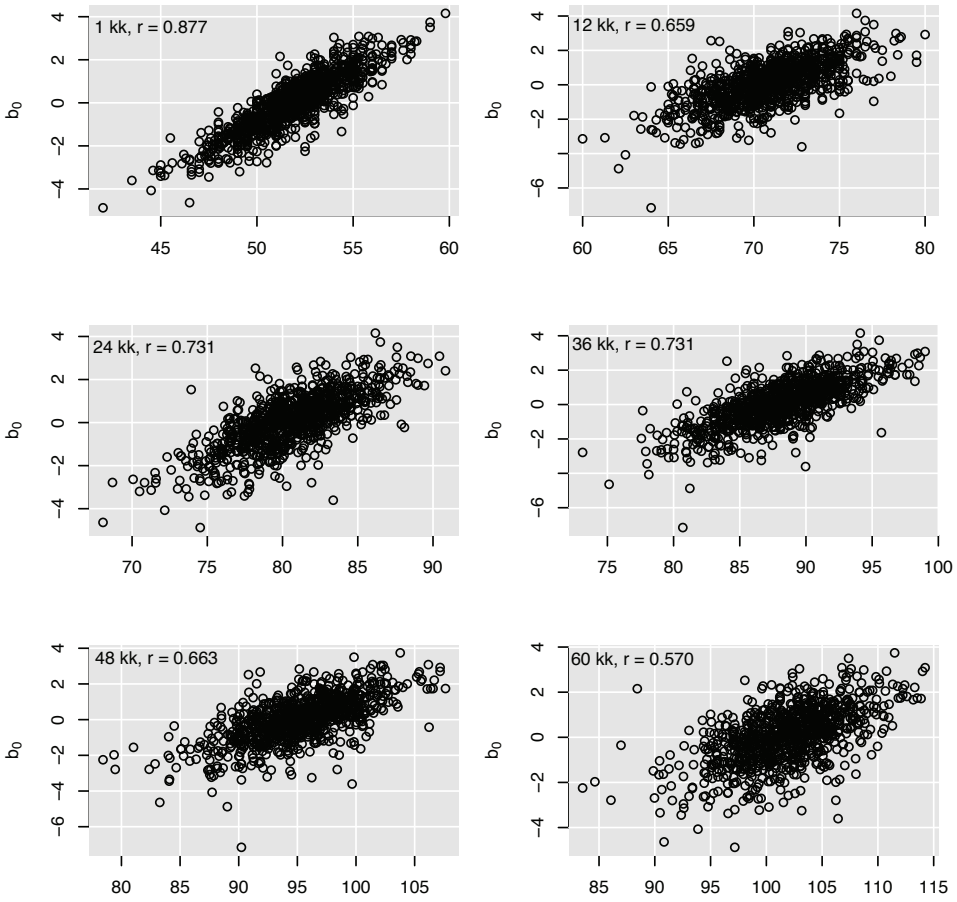
Kuvion 4.6 regressiokäyrissä on nähtävissä erilaisia asetelmia. Se, missä sekamalli onnistuu odotusten mukaisesti, on kahden pituusmittauksen sisältävän yksilön kasvukäyrän muodostamisessa. Esimerkkinä on yksilö C. Tavallinen lineaarinen regressio tuottaisi kahden pisteen välille suoran, mutta lineaarisen sekamallin tilanteessa mainitun yksilön kasvukäyrä on populaation kasvukäyrää mukaileva yksilöllisellä erolla. Näin kasvulle luodaan oletettu käyrä iän mukaan. Lisäksi yksilöllä B voidaan huomata kasvukäyrältä huomattavasti poikkeava havainto, jota ei aineiston puhdistuksen yhteydessä ole määritelty mittausvirheeksi. Suhteessa yksi-



Kuvio 4.6. Muutamien yksilöiden mallinnetut kasvukäyrät *(musta viiva)*, havaitut mittaukset *(ympyrät)* ja kiinteiden vaikutusten keskiarvokäyrä *(harmaa viiva)*. X-akselilla ikä kuukausina ja y-akselilla pituus senttimetreinä.

lön mallin parametreihin valitulla raja-arvolla nämä poikkeavat havainnot saatetaan merkitä mittausvirheiksi ja poistaa puhdistetusta aineistosta. Yksilöt A ja E kuvaavat tilannetta, jossa puuttuvia havaintoja on pitkittäistutkimuksen alku- tai loppupäässä. Myös yksilöiden D ja F tilanteessa mallinnus onnistuu hyvin.

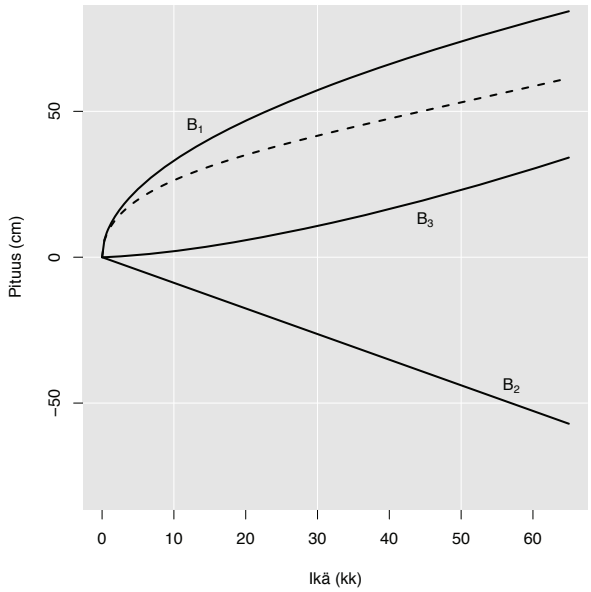
Mallin satunnaisvaikutuksista b_0 korreloi voimakkaasti ($r = 0.877$) yhden kuukauden iässä tehdyn pituusmittauksen kanssa, mutta korrelaatio heikkenee myöhemmissä pituusmittauksissa, minkä voi nähdä kuviosta 4.7. Luonnollisesti satunnaisvaikutus b_0 on yhteydessä yksilön pituuden aloitustasoon, mutta kuvio havainnollistaa, että syntymäpituudella on osittaista yhteyttä pituuteen vielä viiden vuoden iässä. Toisaalta pituuskasvu voi hidastua tai nopeutua myöhemmin. Satunnaisvaikutuksilla b_1, b_2 ja b_3 ei ole huomattavaa korrelaatiota eri iässä tehtyjen pituusmittauksien kanssa.



Kuvio 4.7. Mallin (4.2) satunnaisvaikutuksen b_0 korrelaatio pituuden kanssa eri mittausajankohtina.

Lineaarisen mallin tulkintaa ei voida vain selittävän muuttujan muunnoksia käyttäessä pelkistää yksittäisten parametrien analysoimiseksi, vaan näitä satunnaisvaikutuksia on hyvä katsoa enemmänkin yhdessä kuin erikseen. Toisaalta näiden yksittäisten muuttujien välisten korrelaatioiden avulla joitain yhteyksiä voidaan näh-

dä. Kuitenkaan iän muunnosten yksittäisten parametrien vertailu kasvukäyrään ei tuo esille merkittäviä tulkintoja aineistosta, sillä iän muunnosten parametrien vaikutukset kohdistuvat pituuteen huomattavan eri lailla, kuten kuviosta 4.8 nähdään. Yksittäisten tunnuslukujen luomiseksi voidaan ottaa käyttöön pääkomponentit.



Kuvio 4.8. Mallin (4.2) kiinteiden vaikutusten kantafunktioiden suhde ikään, kun $\hat{\beta}_0 = 0$, missä käyrät $B_1 = \hat{\beta}_1 \times \sqrt{\text{ikä}}$, $B_2 = \hat{\beta}_2 \times \text{ikä}$, $B_3 = \hat{\beta}_3 \times \text{ikä}^{\frac{3}{2}}$ ja katkoviivalla on kiinteiden vaikutusten yhteisvaikutus $B_1 + B_2 + B_3$.

4.4 Satunnaisvaikutusten vertailu

Valitussa mallissa on huomattava iän muunnoksia vastaavien satunnaisparametrien välinen korrelaatio ($|r| > 0.87$). Tällöin satunnaisparametrit sisältävät jotakuinkin saman informaation aineistosta. Mallia voidaan pitää yliparametrisoituna, mutta iän muunnosten käyttö mallinnuksessa on tarpeellista iän ollessa ainoa pituutta selittävä tekijä.

Eräs keino parametrien vähentämiseen on pääkomponenttianalyysin käyttö, jolloin samankaltaisten muuttujien informaatio pyritään kokoamaan yhteen tai useampaan muuttujien yhdistelmään, pääkomponenttiin. Tutkielman sekamallin (4.2) tuottamista yksilöiden satunnaisvaikutuksista on hyvä muodostaa pääkomponentteja, sillä siten yksilöiden välisten erojen tulkittavuus on parannettavissa, kun kiinteiden vaikutusten parametrit tarjoavat vain muuttujien lähtötason. Lisäksi satunnaisvaikutusten pääkomponenttianalyysissa voidaan käyttää mallin antamia yksilöiden satunnaisvaikutuksia, kun kiinteiden vaikutusten kohdalla on käytettävä koko havaintoaineiston mittauksia.

Pääkomponenttianalyysin avulla satunnaisvaikutusten iän muunnosten parametrit b_1, b_2 ja b_3 mallista (4.2) voidaan tiivistää yhteen satunnaiskomponenttiin (pk), joka edustaa suurimman vaihtelun suuntaa satunnaiskomponenttien muodostamassa

avaruudessa. Aineistona on tällöin yksilöiden satunnaisvaikutusten matriisi (1230×3). Pääkomponenttianalyysissa käytettäville muuttujille on tarpeellista tehdä standardointi, koska muuttujien hajonnat ovat hyvin erilaisia, kuten sivulla 23 on listattu.

Satunnaisvaikutusten standardointiin käytetään muuttujien arvoja $\text{sd}(b_1) = 1.16$, $\text{sd}(b_2) = 0.35$ ja $\text{sd}(b_3) = 0.042$. Mallin (4.2) tuottaminen klustereiden standardoitujen satunnaisvaikutusten matriisin satunnaiskomponenteista muodostetaan ensimmäinen pääkomponentti. Pääkomponenttipistemääräksi i. lapselle tämän omia satunnaisvaikutuksia käyttäen muodostuu

$$\text{pk}_i = 0.564 \times \frac{b_{1i}}{\text{sd}(b_1)} - 0.588 \times \frac{b_{2i}}{\text{sd}(b_2)} + 0.579 \times \frac{b_{3i}}{\text{sd}(b_3)}.$$

Ensimmäinen pääkomponentti selittää standardoitujen satunnaisparametrien kokonaisvaihtelusta 95.12 prosenttia, kun sitä vastaava ominaisarvo (varianssi) on 2.85. Tämän jälkeen pääkomponenttipistemäärää pk_i voidaan käyttää approksimoimaan i. yksilön kasvukäyrää mallin (4.2) muunnoksella

(4.3)

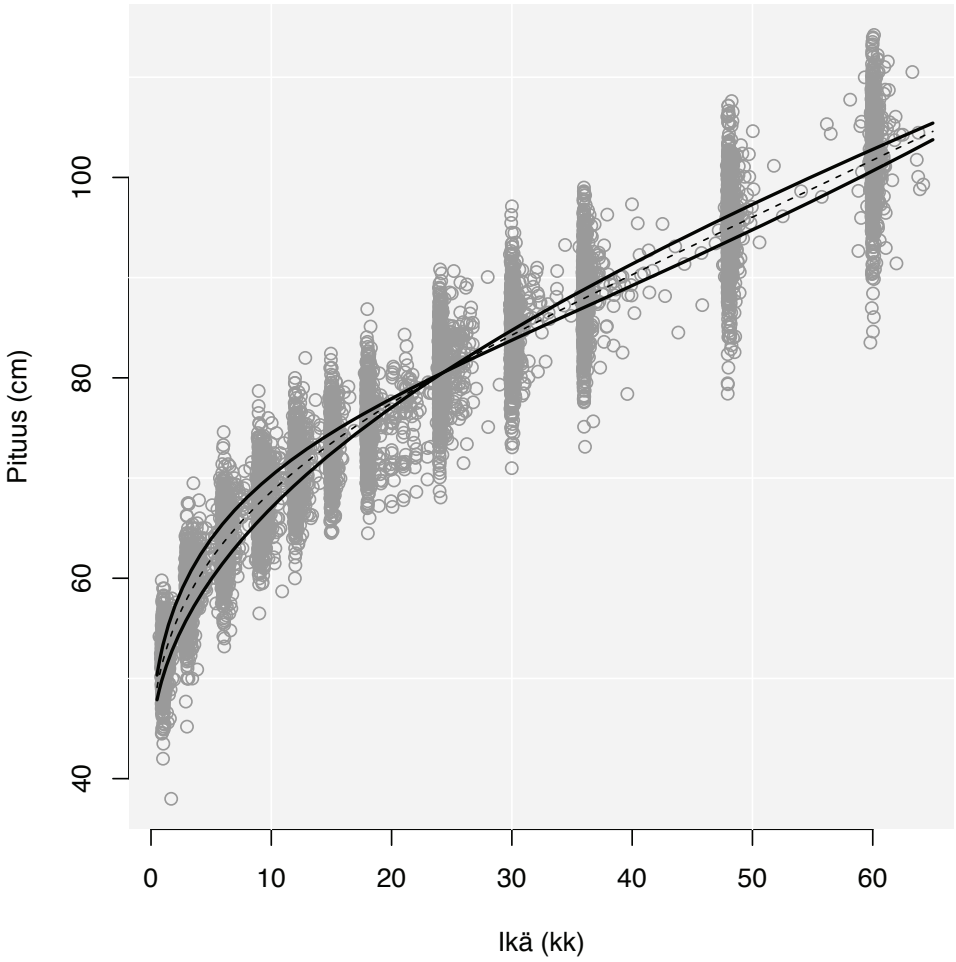
$$\begin{aligned}\text{pituus}_i &= \beta_0 + \beta_1\sqrt{\text{ikä}_i} + \beta_2 \times \text{ikä}_i + \beta_3 \times \text{ikä}_i^{\frac{3}{2}} + b_{0i} \\ &\quad + \text{pk}_i \left[0.564 \times \text{sd}(b_1)\sqrt{\text{ikä}_i} - 0.588 \times \text{sd}(b_2) \times \text{ikä}_i + 0.579 \times \text{sd}(b_3) \times \text{ikä}_i^{\frac{3}{2}}\right].\end{aligned}$$

Tämä on approksimaatio kasvukäyrästä, koska mallissa käytetään vain ensimmäistä pääkomponenttia.

Näin ollen sekamallin parametreina pidetään kiinteinä vaikutuksina alkuperäisiä iän kolmea muunnosta sekä kiinteää ja satunnaista tasovaihtelua, mutta iän satunnaisvaikutukset korvataan niiden yhdistelmällä pk. Tällainen malli ei kuitenkaan selitä aineistoa paremmin kuin malli (4.2). On myös huomattava, että pk ei korreloi pituuden kanssa, kun aineisto jaetaan pienempiin osiin mittauskertojen mukaan kuten aiemmassa kuvailussa b_0:lle tehtiin. Tämä on odotettu tulos, sillä alkuperäisen sekamallin satunnaisvaikutuksista vain b_0 korreloi pituuden kanssa. Lisäksi yksilöiden parametrit b_0 ja pk eivät korreloi keskenään.

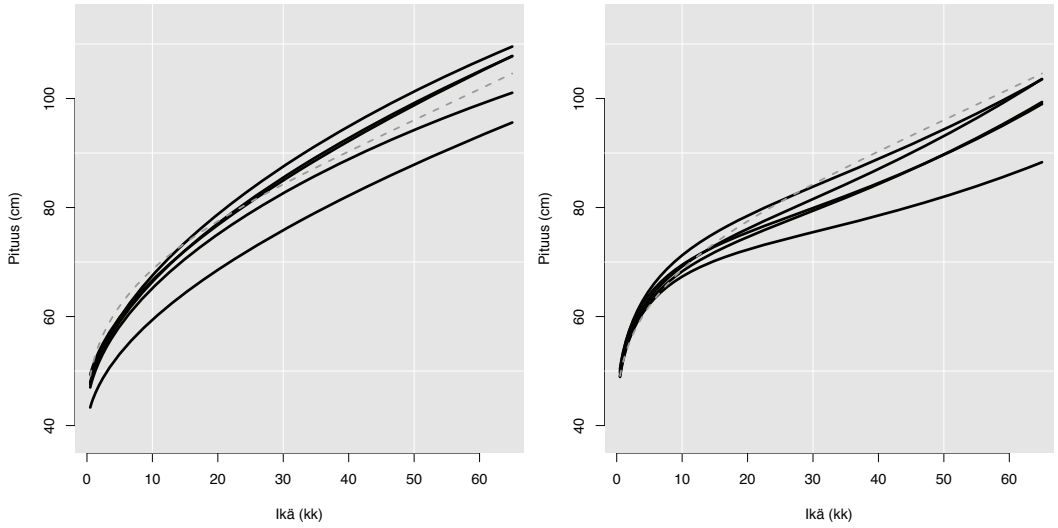
Muodostetusta pääkomponentista voidaan piirtää pituuskasvukäyrälle kahden keskihajonnan ääriarvojen käyrät sijoittamalla pääkomponenttia vastaavan ominaisarvon neliöjuuren arvo kerrottuna kahdella mallin (4.3) parametrin pk_i tilalle. Tämä vaihtelu on kuvattu kuviossa 4.9.

Tällaiset käyrät kuvaavat peruskäyrän vaihtelua siinä, missä vaiheessa kasvu on nopeampaa kuin aineistossa keskimäärin. Joillakin yksilöillä kasvu on nopeampaa alussa, kun toisilla kasvu on nopeimmillaan vasta myöhemmin. Kun kasvun havainnollistamisen täydentämiseen otetaan kuviot 4.4 ja 4.8, voidaan ajatella, että pituuskasvua määrittelevät satunnaisparametrit tasapainottelevat keskenään. Tämä voidaan nähdä satunnaisvaikutusten b_1 ja b_2 korrelaatiosta, joka on voimakkaasti negatiivinen näiden satunnaisvaikutusten vaikuttaessa pituuteen erisuuntaisesti iän funktiona.



Kuvio 4.9. Parametrin *pk* kahden hajonnan äärirajat on piirretty yhtenäisellä viivalla katkoviivalla esitetyn mallin (4.2) keskiarvokäyrän ympärillä.

Parametrin pk vaihteluväli on $(-5.68, 6.73)$ ja keskihajonta on 1.69. Suuria ja pieniä pk-arvoja saavien yksilöiden kasvukäyriä mallilla (4.2) on piirretty kuviossa 4.10. Tällä tavalla parametri pk asettaa yksilöt järjestykseen ja erityisesti suuria pk:n arvoja saavien yksilöiden kasvukäyrät käyttäytyvät niin, että alun nopean kasvun jälkeen yhden ja neljän ikävuoden välillä kasvu on hitaampaa. Pienen pk:n saavat yksilöt kasvavat jotakuinkin nopeammin yhden ikävuoden kohdalla. Siten pääkomponenttipistemäärä pk tarjoaa erilaisen näkökulman kasvukäyrien tarkasteluun.



Kuvio 4.10. Pienimmät *(vas.)* ja suurimmat *(oik.)* pk:n arvot saavien yksilöiden kasvukäyrät mallin (4.2) tuottamien parametrein esitettynä. Katkoviivalla mallin (4.2) keskiarvokäyrä.

Yhden pääkomponentin sekamallia (4.3) ei siis kannata pitää alkuperäisen mallin (4.2) vertaisena, mutta sen pääkomponenttipistemäärät antavat lisää tulkintamahdollisuuksia. Tämä malli antaa aineistolle karkean approksimaation ja lisäksi ominaisuutensa pk suhteen se auttaa erottelemaan yksilöitä.

4.5 Kasvukäyrien ryhmittely

Tutkielman aineistossa on 1230 yksilöä, jotka käytettävissä olevilla taustamuuttujilla voidaan jakaa sukupuolittain kahteen ryhmään. Kun tilastolliselta näkökannalta mallinnetaan pituuskasvua, sukupuoli ei ole välttämätön jakaja. Mallin (4.2) satunnaisvaikutukset sisältävät informaation sukupuolten mahdollisesta erilaisuudesta, mutta ryhmittelyanalyysilla yksilöt voidaan jakaa useampiin tai havainnoillistavimpiin ryhmiin. Tällöin on kuitenkin syytä testata, ettei näin saatavien ryhmien sukupuolijakauma ole vääristynyt. Luonnollinen oletus on, että tutkielman aineisto voidaan jakaa kahteen ryhmään sukupuolen mukaan, mutta oletettavasti kasvukäyrät jakaantuvat useampiin erilaisiin ryhmiin.

Koska ryhmittelyjä voidaan tehdä monella eri tapaa, niin ryhmien määrän kuin käytetyn menetelmän mukaan, on syytä tarkastella tuloksia taulukoiden ja puuku-

vioiden avulla. Ryhmittelyssä analysoidaan mallin (4.2) satunnaisvaikutuksia, mutta niiden hajonnat ovat eri mittaluokkaa, minkä vuoksi ne standardoidaan, jotta suurimman hajonnan sisältävä muuttuja ei vaikuta väärällä tavalla.

Tutkielmassa keskitytään hierarkkisiin ryhmittelymenetelmiin ja etäisyysmittana tutkielmassa käytetään euklidista etäisyyttä. Populaatiota mallinnettaessa pyritään parhaaseen 3–8 ryhmään vertailemalla yhden sidoksen, keskimääräisen sidoksen, täydellisen sidoksen ja Wardin menetelmää sekä näiden puukuvioita.

4.5.1 Ryhmien määrä

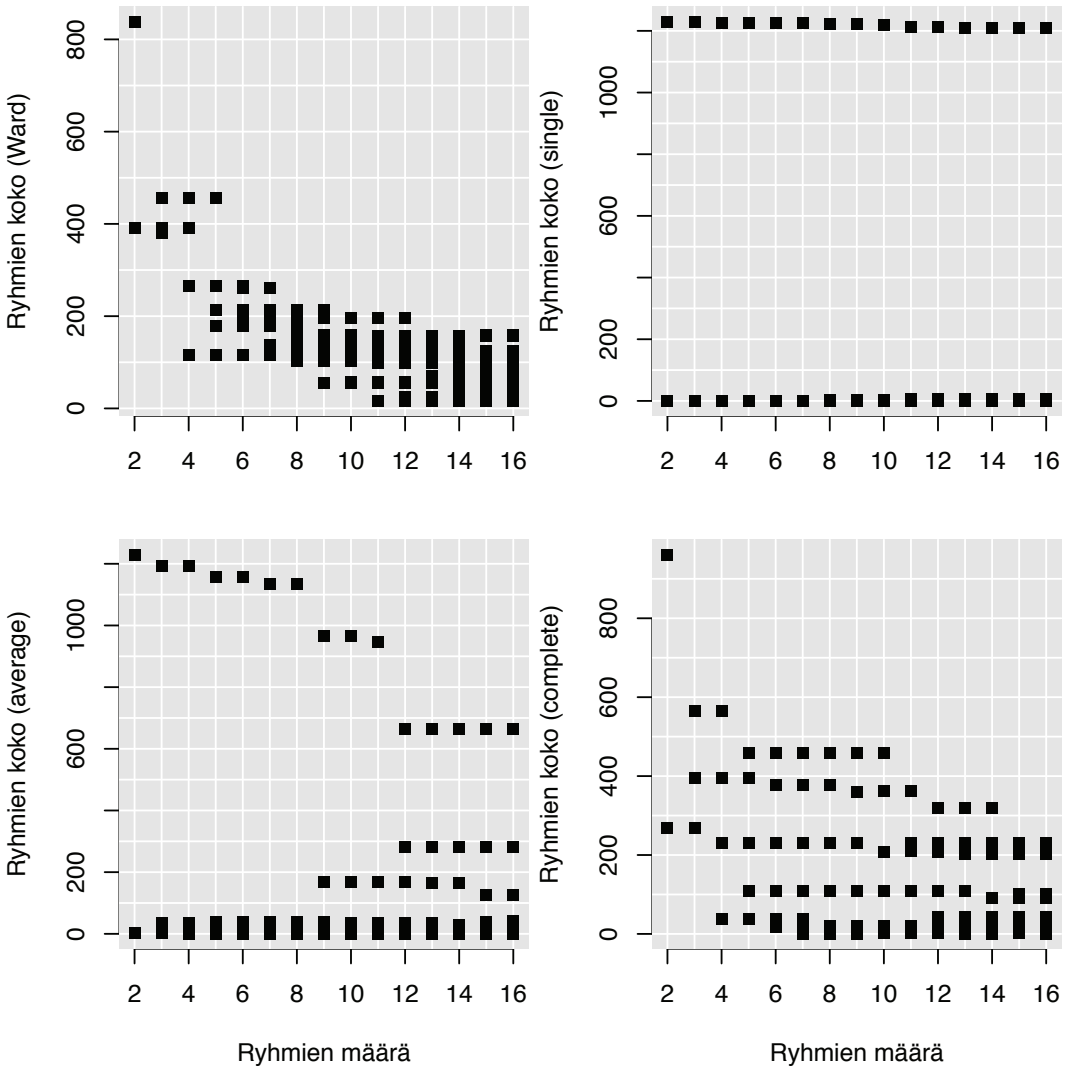
Ryhmittely tehdään mallin (4.2) tuottamien standardoitujen satunnaisvaikutusten b_0, b_1, b_2 ja b_3 perusteella, jolloin lapsen syntymäpituus ja pituuskasvun käyttäytyminen muodostavat kasvukäyrien ryhmittelyn. Näin yksilöllisten kasvukäyrien väliltä pyritään löytämään samankaltaisuuksia.

Oletettavasti aineistossa on poikkeavia kasvukäyriä, jotka ovat verrattain erilaisia kuin muut. Toisilla menetelmillä ryhmiteltäessä tällaiset yksilöt jäävät omiksi tai hyvin pieniksi ryhmikseen. Kun ajatellaan ryhmiteltyjä yksilöitä tutkielman aineiston populaatiossa, kasvukäyrien mallinnuksen kannalta poikkeavien havaintojen poistaminen ei välttämättä ole tarpeellista ja selkeimmät tulokset voidaan saada sillä, että ryhmien havaintomäärät vastaavat toisiaan.

Tutkielman aineistoon suoritettu ryhmittelyanalyysi tuottaa ryhmittelymenetelmän mukaan erilaisia ryhmäkokoja. Yhden sidoksen menetelmä ja keskimääräisen sidoksen menetelmä erittelevät ryhmiä yksittäisiksi ryhmiksi vielä silloinkin, kun ryhmien määrä on pieni. Yhden sidoksen menetelmä tekee ryhmittelyn jopa niin, että se erottaa vain muutamia ryhmiä kerrallaan ylimmältä tasolta. Täten yhden sidoksen menetelmä löytää erittäin poikkeavat ryhmät.

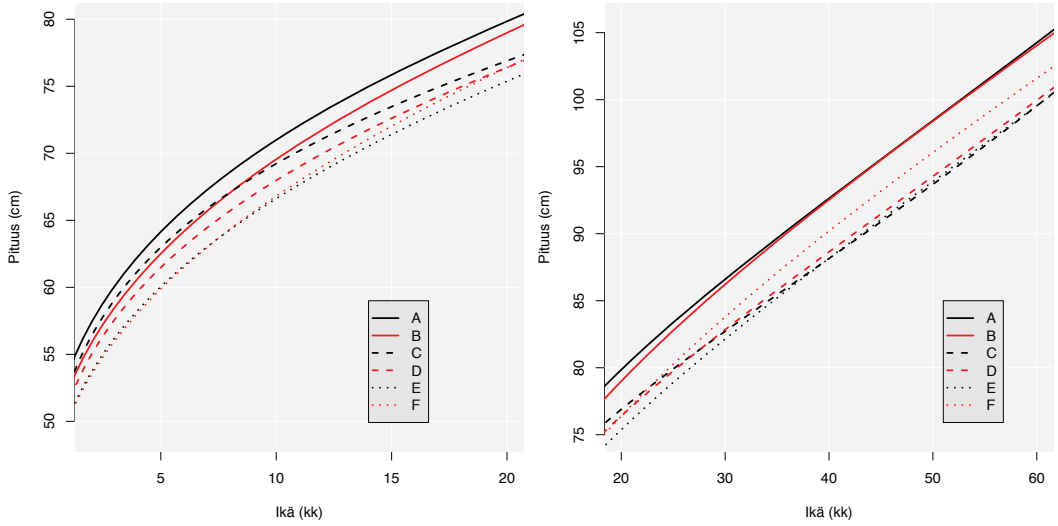
Kuviossa 4.11 on havainnollistettu puukuvion katkaisu ryhmien lukumäärän mukaan neljällä eri ryhmittelymenetelmällä. Mitä lähempänä vasenta alakulmaa tässä kuviossa on pisteitä, sitä herkemmin ryhmittelymenetelmä erottaa pieniä ryhmiä erilleen muista. Näin ollen yksittäisen ja keskimääräisen sidoksen menetelmät erottavat herkemmin poikkeavimmat yksilöt erilleen omiksi ryhmikseen. Wardin menetelmä pyrkii ryhmien kokojen tasapainoisuuteen, kun taas täydellisen sidoksen menetelmä on tavallaan näiden ääripäiden välimuoto. Esimerkiksi keskimääräisen sidoksen menetelmä pitää hyvin pienenä syntyneen ja keskimääräistä pienemmäksi jäävän lapsen omana ryhmänään viimeiseen ryhmien yhdistämiseen saakka. Yksittäisen sidoksen tapauksessa monet yksilöt yhdistetään muihin ryhmiin vasta ryhmittelyn loppuvaiheissa.

Tasaisten ryhmien saamiseksi ja populaation yleistämiseksi Wardin menetelmä tarjoaa konkreettisen ratkaisun kuuden, seitsemän tai kahdeksan ryhmän mukaan. Kuviossa 4.12 on kuvattu kuuden ryhmän keskikäyrät, mistä nähdään erilaisia kasvun käyttäytymisiä. Täydellisen sidoksen menetelmällä voidaan erottaa kuusi ryhmää ennen kuin yksi osa erotetaan 231 yksilön ryhmästä erilleen seitsemänneksi ryhmäksi. Kuvioon 4.13 on kuvattu kuuden täydellisen sidoksen menetelmällä erotetun ryhmän keskikäyrät. Lisäksi täydellisen sidoksen menetelmällä ryhmiteltyjen muuttujien kaksiulotteista jakaumaa on havainnollistettu kuviossa 4.14, mistä näh-

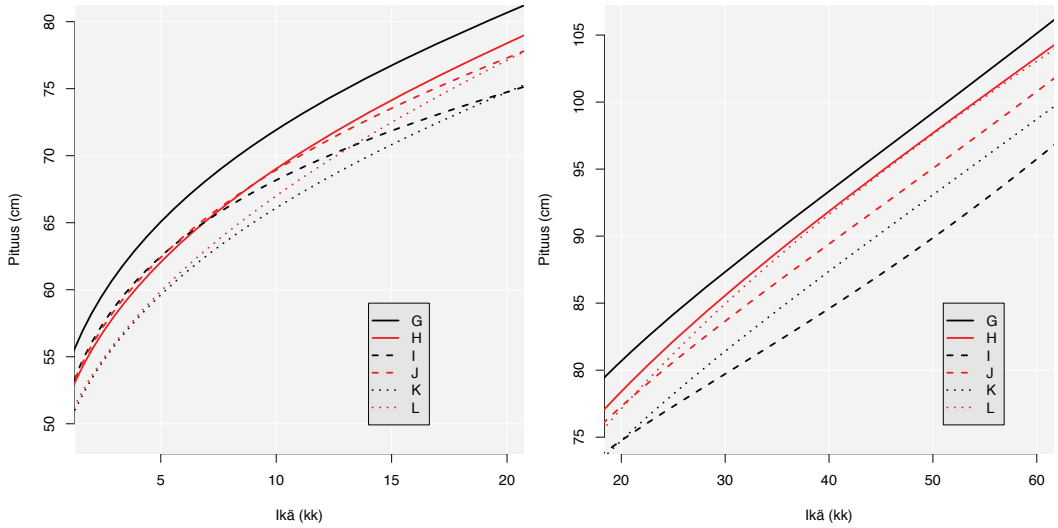


Kuvio 4.11. Hierarkkisten ryhmittelymenetelmien tuottamien ryhmien yksilöiden lukumäärät.

dään ryhmittelyn osittelut. Täydellisen sidoksen ja Wardin menetelmä tuottavat molemmat kuusi ryhmää, jotka valitaan tarkasteluun puukuvioiden ja ryhmien sisältämien havaintojen määrän perusteella.



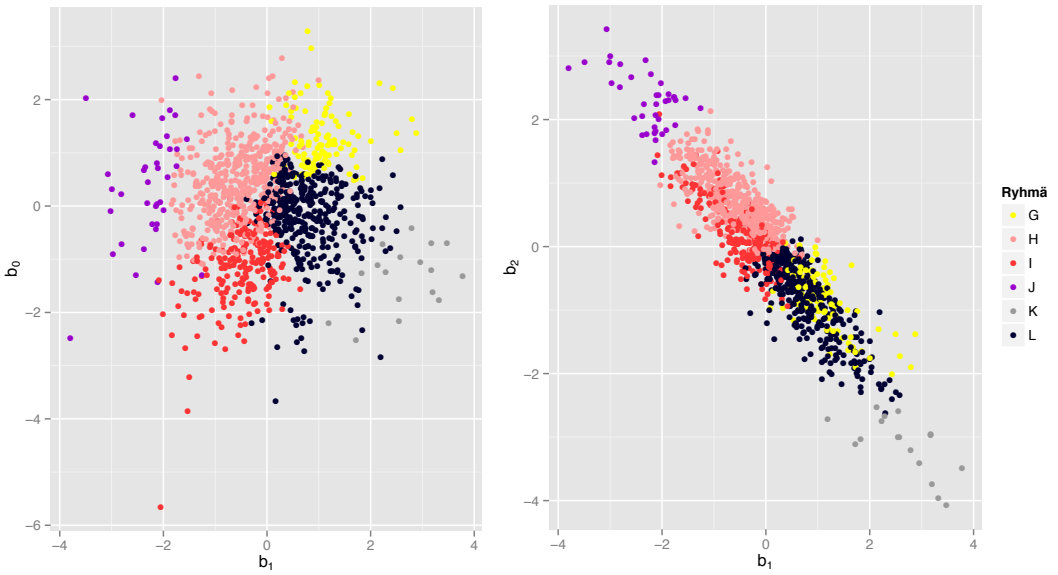
Kuvio 4.12. Kuuden Wardin menetelmällä ryhmitellyn ryhmän keskikäyrät.



Kuvio 4.13. Kuuden täydellisen sidoksen menetelmällä ryhmitellyn ryhmän keskikäyrät.

4.5.2 Ryhmien kuvailu

Näille kahdella tavalla muodostetuille ryhmille lasketaan satunnaisvaikutusten keskiarvot ja niitä vastaavat kasvukäyrät piirretään erilaisten ajan funktiona käyttäytyvien pituuskasvujen ilmentämiseksi (kuviot 4.12 ja 4.13). Menetelmien tuotta-



Kuvio 4.14. Täydellisen sidoksen menetelmällä saatujen kuuden ryhmän standardoitujen parametrien b_0, b_1 ja b_2 sijoittuminen suhteessa toisiinsa.

mien ryhmien käyrillä on yhteneväisyyksiä, mutta selkeänä erona ovat kasvukäyrien suuremmat etäisyydet täydellisen sidoksen menetelmässä. On huomattava, että χ^2-riippumattomuustesti ei hyväksy oletusta ryhmien sukupuolijakauman tasapainoisuudesta. Kuitenkin jokaisessa ryhmässä vähintään 28 prosenttia havainnoista on toista sukupuolta. Seuraavassa ryhmien kasvukäyriä kuvaillaan ryhmittelymenetelmän sisällä.

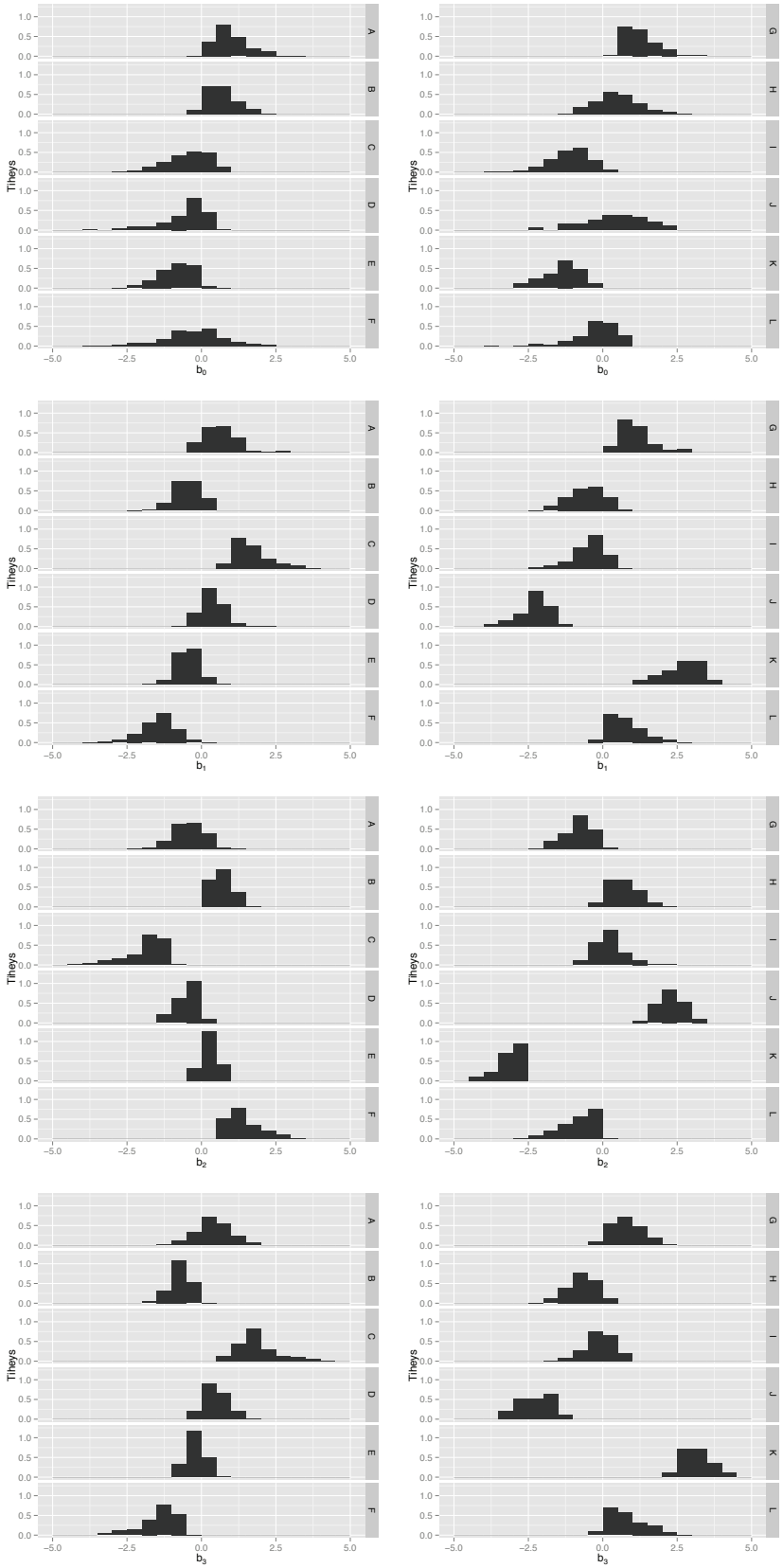
- Wardin menetelmällä valittuja ryhmiä voidaan kuvata seuraavilla tavoilla, kun toista sukupuolta on kasvavassa järjestyksessä 53, 54, 63, 63, 65 ja 67 prosenttia ja näiden ryhmien jäsenten lukumäärän vaihteluväli on 115–266 yksilöä:
  - Korkealta aloittava ja korkealle päätyvä kasvukäyrä, (A).
  - Keskimääräisesti aloittava mutta korkealle päätyvä kasvukäyrä, (B).
  - Keskimääräisesti aloittava mutta voimakkaasti taantuva kasvukäyrä, (C).
  - Keskimääräistä alempaa aloittava mutta taantuva kasvukäyrä, (D).
  - Matalalta aloittava ja matalalle jäävä kasvukäyrä, (E).
  - Matalalta aloittava mutta keskimääräiseksi nouseva kasvukäyrä, (F).
- Täydellisen sidoksen menetelmällä valittuja ryhmiä voidaan kuvata seuraavilla tavoilla, kun toista sukupuolta on kasvavassa järjestyksessä 55, 55, 60, 61, 71 ja 72 prosenttia ja näiden ryhmien jäsenten lukumäärän vaihteluväli on 17–458 yksilöä:
  - Korkealta aloittava ja korkealle päätyvä kasvukäyrä, (G)
  - Keskimääräisesti aloittava mutta korkealle päätyvä kasvukäyrä, (H).

- Keskimääräisesti aloittava mutta voimakkaasti taantuva kasvukäyrä, (I).
- Keskimääräisesti aloittava ja keskimääräiseksi jäävä kasvukäyrä, (J).
- Matalalta aloittava ja matalalle jäävä kasvukäyrä, (K).
- Matalalta aloittava mutta voimakkaasti nouseva kasvukäyrä, (L).

Nämä kaksi menetelmää eivät tuota täysin samanlaisia ryhmiä vaan yksilöt jakautuvat eri tavoin. Voidaan kuitenkin ottaa vertailuun samantapaisia ryhmiä. Esimerkkinä korkealta aloittavat ja korkealle päätyvät ryhmät A ja G. Näiden, samoin kuin muiden, ryhmien parametrien jakaumat on esitetty kuviossa 4.15. Tästä nähdään, että menetelmät löytävät joitakin samankaltaisuuksia. Mielenkiintoisesti samalla tavalla mataliksi kasvukäyriksi kuvailtujen ryhmien E ja K parametrit b_1, b_2 ja b_3 jakaumat sijoittuvat asteikkojen eri kohdille. Tämä selittyy satunnaisvaikutusten korrelaatioilla.

Kun ryhmittelyanalyysissa on mukana satunnnaisvaikutus b_0, tulkinta erottuu tutkielman pääkomponenttianalyysista, jossa tulkitaan vain muuttujia b_1, b_2 ja b_3. Aloitustason sisällyttäminen kasvukäyrätulkintoihin mahdollistaa tasovaihtelun analyysin, mutta rajoittaa pelkkään kasvun kulmakertoimiin keskittyvän pituuskasvun tulkintaa. Toisaalta kumpikin näkökanta, analyysi ilman yksilön kasvukäyrän vakiota tai sen kanssa, tuo erilaisia ulottuvuuksia mallinnukseen ja mahdollistaa niin kasvun tason kuin kasvun kulmakertoimien vertailun.

Pituuskasvukäyrän mallintaminen vakiolla ja iän muunnoksilla tuottaa vaihtelevia malleja mittausten alkaessa 0.5 kuukauden iän kohdalla ja jatkuen noin 60 ikäkuukauteen. Näin pelkkä keskiarvokäyrä tai muutamaan osaan ryhmitellyt kasvukäyrät eivät anna kuin osittaisen, pelkistetyn kuvan malawilaisten lasten pituuskasvun käyttäytymisestä. Yksilötasolta populaation yleistämiseen löytyy monia vaihtoehtoisia ryhmittelyjä, joiden lukumäärän valinnassa tasapainoillaan tarkkuuden monimuotoisuuden ja hallittavuuden yksinkertaisuuden välillä.



Kuvio 4.15. Wardin *(vasen palsta)* ja täydellisen sidoksen *(oikea palsta)* menetelmien tuottamien ryhmien standardoitujen parametrien jakaumat.

5 Johtopäätökset

Sekamallein pystyttiin kuvaamaan lungwenalaisten lasten pituuskasvua hyvin. Sekamallit mahdollistivat myös yksilöiden vertailun. Erityisen hyvää sekamallien käytössä on se, että siten riippuvuuksia sisältävää aineistoa voidaan tarkastella lineaarisen mallin laajennuksena ja epätasapainoinen aineisto on mallinnettavissa. Sekamallien avulla yksilölle saadusta kasvukäyrästä voidaan helposti poimia mittausvirheitä, kun asetetaan haluttu raja-arvo poikkeavuudelle.

Lineaarinen sekamalli pysyi suhteellisen yksinkertaisena, kun pituuskasvumallinnukseen käytettiin lapsen ikää ja sen kahta potenssimuunnosta. Tutkimusasetelma oli sellainen, että vain pituuskasvun mallinnukseen pyrittiin, jolloin huomioitavat muuttujat olivat lapsen ikä ja pituus. Sukupuolen huomioimatta jättäminen ei aiheuta ongelmaa, koska yksilöiden satunnaisvaikutukset muodostavat jokaiselle lapselle onnistuneen pituuskasvumallin sukupuolesta välittämättä.

Mallinnusta on mahdollista tehdä tarkemmaksi sillä uhalla, että sen tulkitseminen vaikeutuu. Esimerkiksi sekamallien regressiosplini-menetelmä *(mixed-effects regression spline, MERS)* lisää sekamalleihin solmukohtia, joissa pituuskasvun mallinnus ositetaan pituuden kasvunopeuden mukaisiin vaiheisiin. Tällä tavalla voisi olla mahdollista kuvata kutakin solmukohtien välistä aluetta ehkä vain yhdellä iän muunnoksella, jolloin selittävien muuttujien voimakas korrelaatio poistuisi. Myös pituuskasvun taantumista tai kasvupyrähdyksiä voidaan mahdollisesti hallita paremmin. Lisäksi pitkittäisaineiston osapopulaatioiden tunnistamiseen on mahdollista käyttää trajektorianalyysia, jolla voidaan korvata ryhmittelyanalyysin tuottamat ryhmäjaot.

Pääkomponenttianalyysin avulla ei voitu parantaa valittua sekamallia, mutta pääkomponenttien käytöllä mallin havainnollistamiseen saadaan uusia näkökulmia. Tällä tavalla keskimääräisen kasvun kulmakertoimien vaihtelua voitiin tulkita. Lisäksi yksilöille pystyttiin antamaan yksi arvo, jonka perusteella kasvun käyttäytymistä pystyttiin jaottelemaan. Jatkotutkimus pääkomponenttien käytöstä lineaarisessa sekamallinnuksessa tarjonnee monia uusia ulottuvuuksia sekamallien tuottamiseen. Useamman muuttujan aineistossa pääkomponenttipistemääriä voidaan käyttää mallinnuksessa ja ryhmittelyssä.

Ryhmittelyanalyysi tarjoaa mielenkiintoisia mahdollisuuksia kuvata aineistoa ja sitä voidaan käyttää monin tavoin. Poikkeavien yksilöiden löytämisestä on hyötyä, kun paikannetaan mittausvirheitä ja poistetaan havaintoja aineistosta. Erityisesti korkealla puukuviossa isompiin ryhmiin liittyvät yksittäiset yksilöt osoittavat poikkeuksellisuutta populaatiosta. Toisaalta erittäin poikkeavat ryhmät saatetaan ottaa tutkittavaksi muista erillään. Joka tapauksessa ryhmittelyanalyysin monet ryhmittelytavat tarjoavat valtavan määrän vertailua, jonka avulla selkeimmät tai informaatioltaan tärkeimmät kasvukäyrät voidaan erottaa. Jo itsessään kasvukäyrien mahdollisimman tarkka ryhmittely esimerkiksi eri alkutasot tai taustamuuttujat huomioiden mahdollistaisi laaja-alaisen tutkimuksen.

Tutkielmassa käsiteltiin muutamia menetelmiä kasvukäyrien mallintamiseen, hal-

lintaan ja tutkimiseen. Suurelta osin tarkastelut olivat visuaalisia, minkä vuoksi menetelmien käyttöä on seuraavaksi analysoitava myös tilastollisin testein. Satunnaisvaikutusten avulla voidaan mallintaa monenlaista yksilöllistä vaihtelua. Kun aineiston mahdollisuuksia on esitetty laaja-alaisesti, on helpompi tarkentaa tulevat tutkimuskysymykset mielenkiintoisiksi nousseisiin seikkoihin. Aineiston havainnollistaminen on ensimmäinen askel aineiston hallintaan, mikä vaatii aineiston oikeiden ulottuvuuksien löytämistä.

Terveyden tutkimusprojekteissa osallistujien taustamuuttujien jakaumat vaikuttavat tutkimustuloksiin joskus hyvinkin merkittävästi. Tässä tutkielmassa taustamuuttujia, kuten äidin koulutus ja aikaisemmat synnytykset, ei hyödynnetty vaan mallinnus pohjautui lapsen ikään huomioimatta ennenaikaista synnytystä. Nämä tekijät huomioiden kasvukäyrävertailua voidaan laajentaa pienempiä ryhmiä kattaviksi vertailuiksi, jolloin populaatiosta on mahdollista löytää syitä eri kasvukäyrien muodolle. Ravinnon vaikutus pituuskasvun taantumiseen ja kasvupyrähdyksiin vaatii tietoa ruuan saatavuudesta ja sen nauttimisesta.

LAIS-tutkimus ja muut aliravitsemuksen ehkäisyyn perustuvat tutkimukset, joissa Tampereen yliopiston kansainvälisen lääketieteen yksikkö on mukana, ovat äärimmäisen mielenkiintoisia ja monivivahteisia projekteja. Niin tutkimuskysymysten ja aineistonkeruun kuin tilastollisten merkitysten ja muuttujien tulkinnan osalta vaiheita on monia. Näin ollen tämä tutkielma on vain pieni osa siitä, mihin kaikkeen keräysaineistoilla on mahdollisuus päästä. Jatkotutkintojen suorittaminen ja kansainvälinen yhteistyö tuottavat monia uusia näkökulmia tutkimuskentälle. Vaikka tieteellinen tutkimus vaatii tilastollisten menetelmien käyttöä ja mahdollistaa niiden laajennuksen, huomattavin vaikutus voidaan toivottavasti nähdä ihmisten pidempänä ja terveempänä elämänä.

Lähteet

Bates, D. (2005), "Fitting linear mixed models in R", *R News 5(1), 27–30*.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2013), "lme4: Linear mixed-effects models using Eigen and S4", *R package version 1.0-5*. http://CRAN.R-project.org/package=lme4.

Cameron, N. & Bogin, B. (2012), ”Human Growth and Development”, *Elsevier*.

Cheung, Y. B. (2014), ”Statistical Analysis of Human Growth and Development”, *CRC Press*.

Davison, A. C. (2008), ”Statistical Models”, *Cambridge University Press*.

Demidenko, E. (2004), ”Mixed Models: Theory and Applications", *Wiley Series in Probability and Statistics*.

Fitzmaurice, G. E., Laird, N. M. & Ware, J. H. (2004), ”Applied Longitudinal Analysis", *Wiley Series in Probability and Statistics*.

Hastie, T., Thibshirani, R. & Friedman, J. (2008), ”The Elements of Statistical Learning”, *Springer*.

Henderson, C. R., Kempthorne, O., Searle, S. R. & von Krosigk, C. M. (1959), ”The Estimation of Environmental and Genetic Trends from Records Subject to Culling", *Biometrics, Vol. 15, No. 2 (Jun., 1959), 192–218*.

Jiang, J. (2007), ”Linear and Generalized Linear Mixed Models and Their Applications”, *Springer*.

Johnson, R. A. & Wichern, D. W. (2007), ”Applied Multivariate Statistical Analysis”, *Pearson Prentice Hall*.

Laird, N. M. & Ware, J. H. (1987), ”Random-effectd models for longitudinal data”, *Biometrics 38, 963–974*.

Luntamo, M., Rantala, A.-M., Meshnick, S. R., Cheung, Y. B., Kulmala, T., Maleta, K., & Ashorn, P. (2012), "The Effect of Monthly Sulfadoxine-Pyrimethamine, Alone or with Azithromycin, on PCR-Diagnosed Malaria at Delivery: A Randomized Controlled Trial", *PLoS ONE 7(7): e41123. doi:10.1371/journal.pone.0041123*.

McCulloch,, C. E. & Searle, S. R. (2001), ”Generalized, Linear, and Mixed Models”, *Wiley Series in Probability and Statistics*.

Nissinen, K. (2009), ”Small Area Estimation with Linear Mixed Models from Unit-level Panel and Rotating Panel Data”, *väitöskirja. Jyväskylän yliopisto*.

Pinheiro, J. C. & Bates, D. M. (2000), "Mixed-Effects Models in S and S-PLUS", *Springer*.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & the R Development Core Team (2013), "nlme: Linear and Nonlinear Mixed Effects Models", *R package version 3.1-113*.

R Core Team (2013), R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. http://www.R-project.org/.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992), " Variance Components", *Wiley*.

Unicef (2013), "Malnutrition Status". http://www.childinfo.org/malnutrition_status.html. Viitattu 8.1.2014.

Liite: Mallinnustulos

```
Linear mixed-effects model fit by REML
 Data: lais
       AIC      BIC    logLik
  49800.08 49918.95 -24884.04

Random effects:
 Formula: ~1 + sqrt(Age) + Age + I(Age^(3/2)) | Participant
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
(Intercept)  2.01806874 (Intr) sqr(A) Age
sqrt(Age)    1.91021212 -0.494
Age          0.55078590  0.488 -0.947
I(Age^(3/2)) 0.04209925 -0.470  0.904 -0.987
Residual     1.40180210

Correlation Structure: Exponential spatial correlation
 Formula: ~1 | Participant
 Parameter estimate(s):
    range
0.6012151
Fixed effects: Height ~ sqrt(Age) + Age + I(Age^(3/2))
                Value  Std.Error    DF  t-value p-value
(Intercept)  42.11698 0.11855215 11220 355.2613       0
sqrt(Age)    10.45532 0.10924928 11220  95.7015       0
Age          -0.86103 0.02922569 11220 -29.4616       0
I(Age^(3/2))  0.06517 0.00222919 11220  29.2338       0
 Correlation:
             (Intr) sqr(A) Age
sqrt(Age)    -0.835
Age           0.770 -0.969
I(Age^(3/2)) -0.720  0.930 -0.988

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-7.2540912 -0.4814338 -0.0040832  0.4814693  5.7340288

Number of Observations: 12453
Number of Groups: 1230
```